On Conditions for Linearity of Optimal Estimation

Emrah Akyol, *Student Member, IEEE*, Kumar B. Viswanatha, *Student Member, IEEE*, and
Kenneth Rose, *Fellow, IEEE*

Abstract—When is optimal estimation linear? It is well known that when a Gaussian source is contaminated with Gaussian noise, a linear estimator minimizes the mean square estimation error. This paper analyzes, more generally, the conditions for linearity of optimal estimators. Given a noise (or source) distribution, and a specified signal-to-noise ratio (SNR), we derive conditions for existence and uniqueness of a source (or noise) distribution for which the L_p optimal estimator is linear. We then show that if the noise and source variances are equal, then the matching source must be distributed identically to the noise. Moreover, we prove that the Gaussian source-channel pair is unique in the sense that it is the only source-channel pair for which the mean square error (MSE) optimal estimator is linear at more than one SNR values. Furthermore, we show the asymptotic linearity of MSE optimal estimators for low SNR if the channel is Gaussian regardless of the source and, vice versa, for high SNR if the source is Gaussian regardless of the channel. The extension to the vector case is also considered where besides the conditions inherited from the scalar case, additional constraints must be satisfied to ensure linearity of the optimal estimator.

Index Terms—Linear estimation, optimal estimation.

I. INTRODUCTION

CONSIDER a basic problem in estimation theory, namely, source estimation from a signal received through a channel with additive noise, given the statistics of both source and channel. The optimal estimator that minimizes the mean square error (MSE) is usually a nonlinear function of the observation. A frequently exploited result in estimation theory concerns the special case of Gaussian source and Gaussian noise, a case in which the MSE optimal estimator is guaranteed to be linear. An open follow-up question considers the existence of other cases exhibiting such a "coincidence," and more generally the characterization of conditions for linearity of optimal estimators for general distortion measures.

This problem also has practical importance beyond theoretical interest, mainly due to significant complexity issues in both design and operation of estimators. Specifically, the optimal estimator generally involves entire probability distributions, whereas linear estimators require only up to second-order statistics for their design. Moreover, unlike the optimal estimator which can be an arbitrarily complex function that is difficult to implement, the linear estimator consists of a simple matrix-vector operation. Hence, linear estimators are more prevalent in practice, despite their suboptimal performance in general. They also represent a significant temptation to "assume" that processes are Gaussian, sometimes despite overwhelming evidence to the contrary. Results in this paper identify the cases where a linear estimator is optimal, and when the use of linear estimators is justified in practice without recourse to complexity arguments.

The estimation problem in general has been studied intensively in the literature [1]–[6]. Our preliminary results appeared in [7] and [8]. It is known that, for stable distributions[1] (which include the Gaussian distribution as the only finite variance member), the optimal estimator is linear at all signal-to-noise ratios (SNRs). Stable distributions are a subset of a family called infinitely divisible distributions which, as we show in this paper, satisfy the derived necessary conditions for the existence of a matching source/noise distribution such that the optimal estimator is linear at any SNR level. Our main contribution relative to prior work, which studied linearity as it applies simultaneously at all SNR levels, focuses on the linearity of optimal estimation for the L_p norm and its dependence on the SNR level. Specifically, we present the optimality conditions for linearity of optimal estimators at a specified SNR, where optimality is in the sense of the L_p norm. As an important special case, we investigate the $p = 2$ case (MSE) in detail. Note that a similar problem has been studied in [9] and [10] for the special case of the MSE, albeit without further study related to questions of existence and uniqueness of "matching" distributions. We show that the necessary conditions presented in [9] and [10] are subsumed in our general necessary and sufficient conditions, and specify conditions for which such matching distributions exist and are unique. The analysis is then extended to the case of vector spaces. Interestingly, this extension is nontrivial and new constraints, beyond those inherited from the scalar case, must be satisfied to ensure linearity of optimal estimation.

Five results are provided on the linearity of optimal estimation. First, we show that if a given noise (alternatively, a given source) distribution satisfies certain conditions, there always exists a matching source (alternatively, noise) distribution of a given power, for which the optimal estimator is linear. We further identify conditions under which such a matching distribution does *not* exist. Second, we show that if the source and

[1]A distribution is called stable if for independent identically distributed X_1, X_2, X, for any constants a and b, the random variable $aX_1 + bX_2$ has the same distribution as $cX + d$ for some constants c and d [5].

the noise have the same variance, they *must* be identically distributed to ensure the linearity of the optimal estimator. Having established more general conditions for linearity of optimal estimation, one wonders in what precise sense the Gaussian case may be special. This question is answered by the third result. We consider the optimality of linear estimation at multiple SNR values. Let random variables X and Z be source and noise, respectively, and allow for scaling of either to produce varying levels of SNR. We show that if the optimal estimator is linear at more than one SNR value, then both the source X and the noise Z must be Gaussian. In other words, the Gaussian source–noise pair is unique in the sense that it offers linearity of optimal estimators at multiple SNR values (in fact the optimal estimator is linear at all SNR as is well known). As a fourth result, we show that the MSE optimal estimator converges to a linear estimator for any source and Gaussian noise at asymptotically low SNR, and vice versa, for any noise and Gaussian source at asymptotically high SNR.

Finally, we analyze the vector case, where conditions for linearity of optimal estimation are more stringent. We show that for a vector source–channel pair with identical dimensions, the conditions derived for the scalar case become necessary conditions in a transform domain, where the transform jointly diagonalizes the source and channel covariance matrices. We further derive the additional complementary conditions that must be satisfied to achieve sufficiency.

This paper is organized as follows. We review optimal and linear estimation in Section II, present the main result in Section III, its main corollaries in Section IV, the vector case in Section V, and conclusions in Section VI.

II. REVIEW OF OPTIMAL AND LINEAR ESTIMATION

A. Preliminaries and Notation

Let $\mathbb{R}$, $\mathbb{R}^+$, and $\mathbb{N}$ denote the respective sets of real numbers, positive real numbers, and natural numbers. In general, lowercase letters (e.g., x) denote scalars, boldface lowercase (e.g., $\boldsymbol{x}$) vectors, uppercase (e.g., U, X) matrices and random variables, and boldface uppercase (e.g., $\boldsymbol{X}$) random vectors. Unless otherwise specified, vectors and random vectors have length m, and matrices have size $m \times m$. The kth element of vector $\boldsymbol{x}$ is denoted by $[\boldsymbol{x}]_k$ and the (i, j)th element and the kth column of the matrix U by $[U]_{ij}$ and $[U]_k$ respectively. U^{-T} denotes $(U^T)^{-1}$. $\mathbb{E}[\cdot]$, R_X, and R_{XZ} denote the expectation, covariance of $\boldsymbol{X}$ and cross covariance of $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively. ∇ denotes the gradient and $\nabla_{\boldsymbol{x}}$ denotes the partial gradient with respect to $\boldsymbol{x}$. $F^{(k)}(\cdot)$ denotes the kth order derivative of the function $F(\cdot)$, i.e., $F^{(k)}(x) = \frac{d^k F(x)}{dx^k}$.

We consider the problem of estimating source X given the observation $Y = X + Z$, where X and Z are independent, as shown in Fig. 1. Let X and Z be scalar zero mean[2] random variables with respective densities $f_X(\cdot)$ and $f_Z(\cdot)$ and characteristic functions $F_X(\omega)$ and $F_Z(\omega)$. A density $f(x)$ is said to
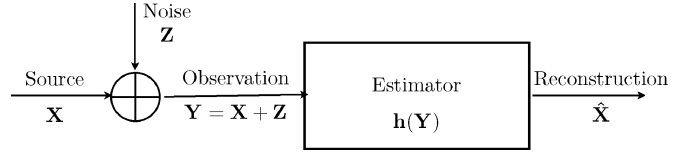

Fig. 1. General setup of the problem.

be symmetric if it is an even characteristic function[3]: $f(x) = f(-x) \; \forall x \in \mathbb{R}$. The SNR is $\gamma = \frac{\sigma_x^2}{\sigma_z^2}$, where $\sigma_x^2 = \mathbb{E}\{X^2\}$ and $\sigma_z^2 = \mathbb{E}\{Z^2\}$. In any statement concerning L_p norm, all random variables are assumed to have finite pth order moments, e.g., in any result associated with MSE we assume finite variances, $\sigma_x^2 < \infty, \sigma_z^2 < \infty$. All the logarithms in the paper are natural logarithms and may in general be complex.

In the rest of this section, we review and derive some preliminary results concerning optimal estimators which will be useful in the following sections in proving our main results. An estimator $h(\cdot)$ is a function of the observation and is said to be optimal if it minimizes the cost functional

$$J(h) = \mathbb{E}\left\{\Phi(X, h(Y))\right\} \tag{1}$$

for a given distortion measure Φ, which is assumed to be first order differentiable. Specializing (1) to a difference distortion measure, we explicitly get

$$J(h) = \int \int \Phi(x - h(y)) f_X(x) f_Z(y - x) dx dy. \tag{2}$$

To obtain the necessary conditions for optimality, we apply the standard method in variational calculus [11]

$$\left. \frac{\partial}{\partial \epsilon} J(h + \epsilon \eta) \right|_{\epsilon=0} = 0 \tag{3}$$

for all variation functions $\eta(\cdot)$. Then, (3) yields

$$\int \int \Phi'(x - h(y)) \eta(y) f_X(x) f_Z(y - x) dx dy = 0 \tag{4}$$

or

$$\mathbb{E}\left\{[\Phi'(X - h(Y)]\eta(Y)\right\} = 0 \tag{5}$$

for all variation functions $\eta(\cdot)$, where Φ' is the derivative of Φ.

B. Optimality Condition for L_p Norm

Hereafter, we will specialize to the case of the L_p metric with $p = 2\rho, \rho \in \mathbb{N}$, i.e., $\Phi(x) = |x|^p$ for even[4] and natural p. Using the fact that $\frac{d}{dx}|x|^p = p\frac{|x|^p}{x}, \forall x \in \mathbb{R} - \{0\}$, we derive the necessary condition for optimality of an estimator as

$$\mathbb{E}\left\{[X - h(Y)]^{p-1}\eta(Y)\right\} = 0. \tag{6}$$

[2]The zero mean assumption is not crucial, but it considerably simplifies the notation. Therefore, it is kept throughout this paper.

[3]Note that this definition requires generalization to symmetry about the mean when one drops the assumption of zero-mean random variables.

[4]Although some of the high level results may be derived for all natural p, in this paper we focus on even p which enables considerable simplification of the results, hence providing much insight and clear intuitive interpretation of the solution.

Note that for $p = 2$, or $\Phi(x) = x^2$, this condition reduces to the well-known orthogonality condition of MSE, i.e., the following holds:

$$\mathbb{E}\{[(X - h(Y)]\eta(Y)\} = 0 \qquad (7)$$

for any $\eta(\cdot)$ function. The MSE optimal estimator $h(Y) = \mathbb{E}\{X|Y\}$ can be directly obtained from (7). The following lemma formally states that the aforementioned necessary condition (6) is also sufficient for minimizing L_p norm.

Lemma 1: The necessary condition stated in (6) is sufficient. Moreover, the optimal estimator is unique almost everywhere (optimal estimators may only differ over a set of zero measure).

Proof: See Appendix A. ∎

C. L_p Optimal Linear Estimation

To derive the optimal linear estimator, the variation function $\eta(y)$ must be made linear to ensure linearity of $h(y) + \epsilon\eta(y)$. Plugging $h(Y) = kY$ and $\eta(Y) = aY$ (for some $a \in \mathbb{R}$) in (6) and omitting straightforward steps, we obtain the condition for optimal linear estimation to be

$$\mathbb{E}\{(X - kY)^{p-1}Y\} = 0. \qquad (8)$$

The optimal scaling coefficient k can be found by plugging $Y = X + Z$ into (8). Observe that for $p = 2$, we get the well-known result $k = \frac{\gamma}{\gamma+1}$.

D. Gaussian Source and Channel

We next consider the special case in which both X and Z are Gaussian, $X \sim \mathcal{N}(0, \sigma_x^2)$, and $Z \sim \mathcal{N}(0, \sigma_z^2)$. The linear estimator

$$h(Y) = \frac{\gamma}{\gamma + 1}Y \qquad (9)$$

is well known to be the optimal MSE estimator. A relatively less known fact is that this linear estimator is optimal more generally for the L_p norm [12]. It is straightforward to show that this linear estimator satisfies (6) by rendering the reconstruction error $X - h(Y)$ independent of Y.

III. CONDITIONS FOR LINEARITY OF OPTIMAL ESTIMATION

In this section, we find the necessary and sufficient conditions in terms of characteristic functions $F_X(\omega)$ and $F_Z(\omega)$ that ensure that $h(Y) = kY$ is the optimal estimator for some $k \in \mathbb{R}$. We first provide the result for the L_p norm, which takes the form of a differential equation that must be satisfied to ensure linearity of optimal estimation, and then specialize it to the MSE case.

A. L_p Norm Condition

As stated previously for any L_p norm result, the characteristic functions of the source and noise $F_X(\omega)$ and $F_Z(\omega)$ are assumed to be pth order differentiable.

Theorem 1: Given an L_p distortion measure, source X and noise Z with characteristic functions $F_X(\omega)$ and $F_Z(\omega)$, respectively, the optimal estimator is linear, $h(Y) = kY$, where

$Y = X + Z$, if and only if the following differential equation is satisfied:

$$\sum_{m=0}^{p-1}\binom{p-1}{m}F_X^{(m)}(\omega)F_Z^{(p-1-m)}(\omega)\left(\frac{k-1}{k}\right)^m = 0. \qquad (10)$$

Proof: See Appendix B. ∎

B. Specializing to MSE: The Matching Condition

In this section, we explore the impact of Theorem 1 for the special case of the MSE distortion metric, i.e., $p = 2$. More precisely, we wish to find the entire set of source and channel distributions such that $h(Y) = \frac{\gamma}{\gamma+1}Y$ is the optimal estimator for a given SNR, γ. Note that this condition was derived, in another context [9], [10], albeit without consideration of important implications which we focus on, including the conditions for existence and uniqueness of matching distributions. Specifically, we identify the conditions for existence and uniqueness of a source distribution that *matches* the noise (and vice versa) in a way that guarantees the linearity of the optimal estimator. We state the main result for MSE in the following theorem.

Theorem 2: Given SNR level γ, and noise Z with characteristic function $F_Z(\omega)$, there exists a source X for which the optimal estimator is linear if and only if the function

$$F(\omega) = F_Z^\gamma(\omega)$$

is a legitimate characteristic function. Moreover, if $F(\omega)$ is legitimate, then it is the characteristic function of the matching source, i.e., $F_X(\omega) = F(\omega)$.

(An equivalent theorem holds where we replace "noise" for "source" everywhere, i.e., given source and SNR level, we have a condition for existence of a matching noise.)

Proof: Plugging $p = 2$ and $k = \frac{\gamma}{\gamma+1}$ in (10) yields

$$\frac{1}{F_X(\omega)}\frac{dF_X(\omega)}{d\omega} = \gamma\frac{1}{F_Z(\omega)}\frac{dF_Z(\omega)}{d\omega} \qquad (11)$$

or more compactly

$$\frac{d}{d\omega}\log F_X(\omega) = \gamma\frac{d}{d\omega}\log F_Z(\omega). \qquad (12)$$

The solution to this differential equation is given by

$$\log F_X(\omega) = \gamma\log F_Z(\omega) + C \qquad (13)$$

where C is a constant. Imposing $F_Z(0) = F_X(0) = 1$, we obtain $C = 0$, which implies

$$F_X(\omega) = F_Z^\gamma(\omega). \qquad (14)$$

∎

Hence, given a noise distribution, the necessary and sufficient condition for the existence of a matching source distribution boils down to the requirement that $F_Z^\gamma(\omega)$ be a valid characteristic function. Moreover, if such a matching source exists, we have a recipe for deriving its distribution.

C. Existence of a Matching Source for a Given Noise

In this section, we study the conditions under which a matching source exists for a given noise distribution. During the course, we also study some important properties relating the matching distributions when they exist.

We begin with Bochner's theorem [3], which states that a continuous function $F : \mathbb{R} \rightarrow \mathbb{C}$ with $F(0) = 1$ is a valid characteristic function if and only if it is *positive semidefinite*.[5] Hence, the existence of a matching source depends on the positive semidefiniteness of $F_Z^\gamma(\omega)$.

We note that characterizing the entire set of $F_Z(\omega)$ where $F_Z^\gamma(\omega)$ is positive semidefinite is a long-standing open problem. Instead, we illustrate the result with various cases of interest where $F_Z^\gamma(\omega)$ is, or is not, positive semidefinite. Let us start with a simple but useful case.

Corollary 1: If SNR $\gamma \in \mathbb{N}$, a matching source distribution exists, regardless of the noise distribution.

Proof: From (14), natural γ implies

$$X = \sum_{i=1}^{\gamma} Z_i \tag{15}$$

where Z_i are independent and distributed identically to Z. Hence, $F_Z^\gamma(\omega)$ is a valid characteristic function and a matching X exists. ∎

Next, we recall the concept of infinite divisibility, which is closely related to the problem at hand.

Definition [13]: A distribution with characteristic function $F(\omega)$ is called infinitely divisible, if for each integer $k \geq 1$, there exists a characteristic function $F_k(\omega)$ such that

$$F(\omega) = F_k^k(\omega). \tag{16}$$

Alternatively, $f_X(\cdot)$ is infinitely divisible if and only if the random variable X can be written for any k as $X = \sum_{i=1}^{k} X_i$ where $\{X_i, i = 1, \dots, k\}$ are independent and identically distributed.

Infinitely divisible distributions have been studied extensively in probability theory [13], [14]. It is known that Poisson, exponential, and geometric distributions as well as the set of stable distributions (which includes the Gaussian distribution) are infinitely divisible. On the other hand, it is easy to see that distributions of discrete random variables with finite alphabets are not infinitely divisible.

Corollary 2: A matching source distribution exists for all $\gamma \in \mathbb{R}^+$ if and only if $f_Z(\cdot)$ is infinitely divisible.

[5]Let $f : \mathbb{R} \rightarrow \mathbb{C}$ be a complex-valued function, and $t_1, \dots, t_s$ be a set of points in $\mathbb{R}$. Then, f is said to be positive semidefinite (nonnegative definite) if for any $t_i \in \mathbb{R}$ and $a_i \in \mathbb{C}$, $i = 1, \dots, s$ we have

$$\sum_{i=1}^{s} \sum_{j=1}^{s} a_i a_j^* f(t_i - t_j) \geq 0$$

where a_j^* is the complex conjugate of a_j. Equivalently, we require that the $s \times s$ matrix constructed with $f(t_i - t_j)$ be positive semidefinite. If function f is positive semidefinite, its Fourier transform is nonnegative everywhere $F(\omega) \geq 0$, $\forall \omega \in \mathbb{R}$. Hence, in the case of our candidate characteristic function, this requirement ensures that the corresponding density is indeed nonnegative everywhere.

Proof: We first note that if $f_Z(\cdot)$ is infinitely divisible, $F_Z^{1/j}(\omega)$ is a valid characteristic function for all natural j, as follows directly from the definition of infinite divisibility. Then, by Corollary 1, it follows that $F_Z^{i/j}(\omega)$ is also a valid characteristic function, which implies that so is $F_Z^r(\omega)$ for all positive rational $r > 0$ since a rational r means that $r = i/j$ for some natural i and j. Using the fact that every $\gamma \in \mathbb{R}^+$ is a limit of a sequence of rational numbers r_n, and by the continuity theorem [5], we conclude that $F_X(\omega) = F_Z^\gamma(\omega)$ is a valid characteristic function, and hence, a matching source exists.

Toward showing the converse, note that if $F_X(\omega) = F_Z^\gamma(\omega)$ is a valid characteristic function for all γ, then $f_Z(\cdot)$ has to be infinitely divisible, because we can always choose $\gamma = \frac{1}{k}$ for $k \in \mathbb{N}$ and set $F_k(\omega) = F_X(\omega)$ in (16). ∎

However, note that at a given SNR, there may exist a matching source, even though $f_Z(\cdot)$ is not infinitely divisible. For example, a finite alphabet discrete random variable V is not infinitely divisible but still can be k-divisible, where $k < |V| - 1$ and $|V|$ is the cardinality of V. Hence, when $\gamma = \frac{1}{k}$, there may exist a matching source, even when the noise distribution is not infinitely divisible. Many examples follow directly from Corollary 1.

We next cite a theorem, regarding analytic characteristic functions, which will be useful in the proofs that follow.

Theorem [13]: A characteristic function $F(\omega)$ is analytic if and only if F has finite moments of all orders and there exists a finite β such that $\mathbb{E}\{|X^k|\} \leq k!\beta^k$, $\forall k \in \mathbb{N}$. This requirement is equivalent to the existence of a moment generating function. A characteristic function $F(\omega)$ is analytic if and only if the moments $\mathbb{E}\{|X^k|\}$ uniquely characterize the distribution, which in general is not the case (see, e.g., [15]).

A useful property of the matching pair, relating to the analyticity of their characteristic functions, is captured by the following corollary.

Corollary 3: If $F_Z(\omega)$ (or $F_X(\omega)$) is analytic, then the matching $F_X(\omega)$ (or $F_Z(\omega)$), if it exists, is analytic.

Proof: Recall the orthogonality property of the MSE optimal estimator (7). Let $\eta(Y) = Y^m$ for $m = 1, 2, 3 \dots M$. Plugging the best linear estimator $h(Y) = \frac{\gamma}{\gamma+1} Y$ and replacing Y with $X + Z$, we obtain the condition

$$\mathbb{E}\left\{ \left[X - \frac{\gamma}{\gamma+1}(X+Z) \right] (X+Z)^m \right\} = 0 \quad \text{for } m = 1, \dots, M. \tag{17}$$

Applying the binomial expansion

$$(X+Z)^m = \sum_{i=0}^{m} \binom{m}{i} X^i Z^{m-i} \tag{18}$$

and rearranging the terms, we obtain M linear equations that recursively relate the $M + 1$ moments of X, i.e., for $m = 1, \dots, M$, we have

$$\mathbb{E}(X^{m+1}) = \gamma \mathbb{E}(Z^{m+1}) + \sum_{i=0}^{m-1} A(\gamma, m, i) \mathbb{E}(Z^{i+1}) \mathbb{E}(X^{m-i}) \tag{19}$$

where $A(\gamma, m, i) = \gamma \binom{m}{i} - \binom{m}{i+1}$.

Note that if $F_Z(\omega)$ is analytic, Z has finite moments of all orders and $\mathbb{E}\{|Z^k|\} \leq k!\beta^k$, $\forall k$. From (19), by induction,

we can show that all moments of X exist and are bounded by $\mathbb{E}\{|X^k|\} \leq k!(\max\{\gamma, 1\}\beta)^k$. This condition is sufficient to show that X also has an analytic characteristic function. ∎

The following corollary identifies a case in which a matching source does not exist.

Corollary 4: For $\gamma \notin \mathbb{N}$, if $F_Z(\omega)$ is real and analytic and it is negative somewhere, i.e., $\exists \omega$ such that $F_Z(\omega) < 0$, then a matching source distribution does not exist.

Proof: We prove this corollary by contradiction. Let $F_Z(\omega)$ be a valid characteristic function. Let us first assume that a matching source, X exists. Hence, from Corollary 3, it follows that X must have an analytic characteristic function $F_X(\omega)$. We will show that this leads to a contradiction. Recall the set of moment (19). It follows by induction over the set of moment equations starting from $m = 1$ that if all odd moments of Z are zero, then so are all odd moments of X. As the noise is symmetric, it follows from analyticity of $F_X(\omega)$ that the matching source must also be symmetric, since moments of X fully characterize its distribution.

However, if $\gamma \notin \mathbb{N}$, by (14), it follows that $F_X(\omega)$ is not real everywhere, and hence, $f_X(\cdot)$ is not symmetric. This contradiction shows that no matching source exists for symmetric noise distributions which are nonpositive semidefinite when $\gamma \notin \mathbb{N}$. ∎

Let us provide a commonly used example distribution to which the aforementioned corollary applies: uniform distribution over $[-a, a]$. In this case, $f_Z(\cdot)$ is symmetric with an analytic characteristic function, but it is not positive semidefinite. The corollary states that, except for natural values of SNR, the optimal estimator is strictly nonlinear for an additive uniform channel. Example 1 illustrates this point numerically.

Remark: As an important application, consider high-resolution quantization theory. Standard high-resolution approximations assume quantization noise independent of (or uncorrelated with) the source [16]. In practice, such approximations can be made explicit by using a dithered quantizer [17] that generates quantization error independent of the source. Then, the quantizer is equivalent to an additive uniform noise channel. The corollary states that, other than for natural values of SNR, a linear decoder (e.g., a Wiener filter at the decoder) is strictly suboptimal for sources encoded at high resolution or by dithered quantization.

D. Uniqueness of a Matching Source for a Given Noise

Note that (14) may have multiple solutions due to multiplicity of complex roots. The following corollary establishes that for a large set of source (or noise) distributions, the matching noise (or source) is unique.

Corollary 5: If $F_Z(\omega)$ (or $F_X(\omega)$) is analytic, then the matching $F_X(\omega)$ (or $F_Z(\omega)$) is unique.

Proof: We prove this corollary from the set of moment (19). Note that every equation introduces a new variable $\mathbb{E}(X^{m+1})$, for $m = 1, \ldots, M$, so each new equation is linearly independent of its predecessors. Let us consider solving these equations recursively, starting from $m = 1$. At each m, we have one unknown ($\mathbb{E}(X^{m+1})$) in a "linear" equation. Since the number of equations is equal to the number of unknowns for each m, and the equations are linear in terms of

the unknown, there must exist a unique moment sequence that solves (19). From Corollary 3, it also follows that X has an analytic characteristic function. Hence, the moment sequence fully characterizes X and the matching source X (if exists) is unique. ∎

IV. IMPLICATIONS OF THE LINEARITY CONDITIONS

In this section, we explore some special cases obtained by varying γ and utilizing the matching conditions for MSE and L_p. We start with a simple but perhaps surprising result.

Theorem 3: Given a source and noise of equal variance, the L_p optimal estimator is linear if and only if the noise and source distributions are identical, i.e., $f_X(x) = f_Z(x), \ \forall x \in \mathbb{R}$ and in which case, the optimal estimator is $h(Y) = \frac{1}{2}Y$.

Proof: For MSE, it is straightforward to see from (14) that, at $\gamma = 1$, the characteristic functions must be identical. Since the characteristic function uniquely determines the distribution [5], $f_X(x) = f_Z(x), \ \forall x \in \mathbb{R}$. In fact, this results apply more generally. This can be observed directly from Theorem 1 that $F_Z(\omega) = F_X(\omega)$ satisfies the necessary and sufficient optimality condition, and hence, this result also applies to the L_p norm distortion measure. ∎

Our next result pertains to the specialty of Gaussian distribution in the context of linearity of optimal estimation. It is well known that linearity of optimal estimation for all SNR levels characterizes the stable family of distributions, which includes Gaussian as the only finite variance member [1], [2], [6], [18], [19]. However, all prior results on characterizing Gaussian density using linearity of optimal estimation consider optimal estimation for *all* SNR levels, $\gamma \in \mathbb{R}^+$.

Let us consider a setup with given source and noise variables which may be scaled to vary the SNR, γ. Can the optimal estimator be linear at multiple values of γ? This question is motivated by the practical setting where γ is not known in advance or may vary (e.g., in the design stage of a communication system). It is well known that the Gaussian source–Gaussian noise pair makes the optimal estimator linear at all γ levels. In the following, we show that this is the only source–channel pair whose optimal estimators are linear at more than one SNR value.

Theorem 4: Let the source or channel variables be scaled to vary the SNR γ. The MSE optimal estimator is linear at two different SNR values γ_1 and γ_2, if and only if source and noise are both Gaussian. Moreover, this claim also holds for L_p norm if the source (or noise) has an analytic characteristic function.

Proof: Let Z_1 and Z_2 denote the noise random variables with variances $\sigma_{z_1}^2, \sigma_{z_2}^2$ and characteristic functions $F_{Z_1}(\omega), F_{Z_2}(\omega)$, respectively. Let us say the noise is scaled by $\alpha \in \mathbb{R}$, i.e., $Z_2 = \alpha Z_1$ and hence $F_{Z_2}(\omega) = F_{Z_1}(\omega\alpha)$ and $\sigma_{z_2}^2 = \alpha^2 \sigma_{z_1}^2$. Let

$$\gamma_1 = \frac{\sigma_x^2}{\sigma_{z_1}^2}, \ \gamma_2 = \frac{\sigma_x^2}{\alpha^2 \sigma_{z_1}^2}. \tag{20}$$

Using (14)

$$F_X(\omega) = F_{Z_1}^{\gamma_1}(\omega), F_X(\omega) = F_{Z_1}^{\gamma_2}(\omega\alpha). \tag{21}$$

Hence

$$F_{Z_1}^{\gamma_1}(\omega) = F_{Z_1}^{\gamma_2}(\omega\alpha). \tag{22}$$

Taking the logarithm on both sides of (22), and applying (20) and rearranging terms, we obtain

$$\alpha^2 = \frac{\log F_{Z_1}(\alpha\omega)}{\log F_{Z_1}(\omega)}. \tag{23}$$

Note that (23) should be satisfied for both α and $-\alpha$ since they yield the same γ. Hence, $F_{Z_1}(\alpha\omega) = F_{Z_1}(-\alpha\omega)$ for all $\alpha \in \mathbb{R}$, which implies $F_{Z_1}(\omega) = F_{Z_1}(-\omega)$, $\forall \omega \in \mathbb{R}$. Using the fact that the characteristic function is conjugate symmetric (i.e., $F_{Z_1}(-\omega) = F_{Z_1}^*(\omega)$), we get $F_{Z_1}(\omega) \in \mathbb{R}$, $\forall \omega$. As $\log F_{Z_1}(\omega)$ is a function from $\mathbb{R} \to \mathbb{C}$, Weierstrass theorem [20] guarantees that there is a sequence of polynomials that uniformly converges to it: $\log F_{Z_1}(\omega) = \sum_{i=0}^{\infty} k_i \omega^i$, where $k_i \in \mathbb{C}$. Hence, by (23) we obtain

$$\alpha^2 = \frac{\sum_{i=0}^{\infty} k_i (\omega\alpha)^i}{\sum_{i=0}^{\infty} k_i \omega^i}, \quad \forall \omega \in \mathbb{R} \tag{24}$$

which is satisfied for all ω only if all coefficients k_i vanish, except for k_2, i.e., $\log F_{Z_1}(\omega) = k_2\omega^2$, or $\log F_{Z_1}(\omega) = 0$ $\forall \omega \in \mathbb{R}$ (the solution $\alpha = 1$ is of no interest). The latter is not a characteristic function, and the former is the Gaussian characteristic function, $F_{Z_1}(\omega) = e^{k_2\omega^2}$, where we use the established fact that $F_{Z_1}(\omega) \in \mathbb{R}$. Since a characteristic function determines the distribution uniquely, the Gaussian source and noise must be the only such pair.

Next, we extend the result to the L_p norm, albeit we require analyticity of the characteristic function of X (or Z_1 and Z_2). Then, due to Corollary 3, matching noises Z_1 and Z_2 also have analytic characteristic functions, and hence, the moments of X, Z_1 and Z_2 are finite (they have moments of all orders) and moments fully characterize the distribution. The extension to L_p requires a different approach. For simplicity, we first derive the result for MSE (now with analyticity imposed) and then extend the arguments to the L_p case. The following relation between the moments of the original and scaled noise should be satisfied:

$$\mathbb{E}(Z_2^m) = \alpha^m \mathbb{E}(Z_1^m) \text{ for } m = 1, \ldots, M+1. \tag{25}$$

Also, a set of moment equations should hold for two SNR values, γ_1 and γ_2. Let us consider the set of moment equations with moments up to M

$$\mathbb{E}(X^{m+1}) = \gamma_j \mathbb{E}(Z_j^{m+1}) + \sum_{i=0}^{m-1} A(\gamma_j, m, i)\mathbb{E}(Z_j^{i+1})\mathbb{E}(X^{m-i}) \tag{26}$$

where $m = 1, \ldots, M$, $j = 1, 2$ and $A(\gamma, m, i) = \gamma\binom{m}{i} - \binom{m}{i+1}$. Similar to the proof of Corollary 5, we note that every equation introduces a new variable $\mathbb{E}(X^{m+1})$, for $m = 1, \ldots, M$, so each new equation is independent of its predecessors. Next, we solve these equations recursively,

starting from $m = 1$. At each m, we have three unknowns ($\mathbb{E}(X^{m+1}), \mathbb{E}(Z_1^{m+1}), \mathbb{E}(Z_2^{m+1})$) that are related "linearly." Since the number of linearly independent equations is equal to the number of unknowns for each m, there must exist a unique solution. We know that the moment sequences of the Gaussian source–channel pair satisfy (26) since it ensures linearity of optimal estimation. The moment sequence of a Gaussian satisfies Carleman's general criterion [15], and therefore, it uniquely determines the corresponding distribution, so the Gaussian source and noise pair is the only solution to (26).

The proof for L_p norm follows the same lines. Note that as mentioned in Section II-D, the same linear estimator is L_p optimal for a Gaussian source–channel pair. Plugging $Y = X + Z$ in the optimality condition with L_p norm (6), we reach a similar set of moment equations. Following similar arguments, we show that this result holds for the L_p norm. ∎

Next, we investigate the asymptotic behavior of optimal estimation at low and high SNR. The results of our asymptotic analysis are of practical importance since they justify the use of linear estimators without recourse to complexity arguments at high and low asymptotic SNR regimes, under certain conditions.

Theorem 5 (For MSE Only): In the limit $\gamma \to 0$, the MSE optimal estimator is asymptotically linear if the channel is Gaussian, regardless of the source. Similarly, as $\gamma \to \infty$, the MSE optimal estimator is asymptotically linear if the source is Gaussian, regardless of the channel.

Proof: We will present a sketch of the proof here, while a more rigorous formal proof is presented in Appendix C. The proof follows from applying the central limit theorem [5] to the matching condition (14). The central limit theorem states that as $\gamma \to \infty$, for any finite variance noise Z, the characteristic function of the matching source $F_Z^\gamma(\omega)$ converges pointwise to the Gaussian characteristic function. Hence, at asymptotically high SNR, any noise distribution is matched by the Gaussian source.

Similarly, as $\gamma \to 0$ and for any $F_X(\omega)$, $F_X^{\frac{1}{\gamma}}(\omega)$ converges pointwise to the Gaussian characteristic function, and hence, the MSE optimal estimator is asymptotically linear if the channel is Gaussian. ∎

Example 1: Let us consider a numerical example that illustrates our findings. Consider a setting where X is Gaussian with unit variance, i.e., $X \sim \mathcal{N}(0, 1)$ and Z is distributed uniformly on the interval $[-a, a]$. Note that this is a typical setting for high rate or dithered quantization of a Gaussian source, in the sense that the quantization error is uniform and independent of the source. We change γ (SNR) by varying a and observe how the optimal estimator ($h(Y) = \mathbb{E}\{X|Y\}$) and associated estimation error ($\mathbb{E}\{(X - h(Y))^2\}$) behave for different γ. We numerically calculated the optimal estimator and the estimation error by sampling the distributions on a uniform grid. Fig. 2 shows how the optimal estimator converges to linear as SNR increases. Note that at $\gamma = 0.1$, optimal estimator is highly nonlinear while at $\gamma = 10$, it practically converges to a linear one. Fig. 3 demonstrates how the estimation error varies with SNR. As theoretically expected (and from Fig. 2), we see a significant difference at $\gamma = 0.1$, which vanishes at high SNRs.
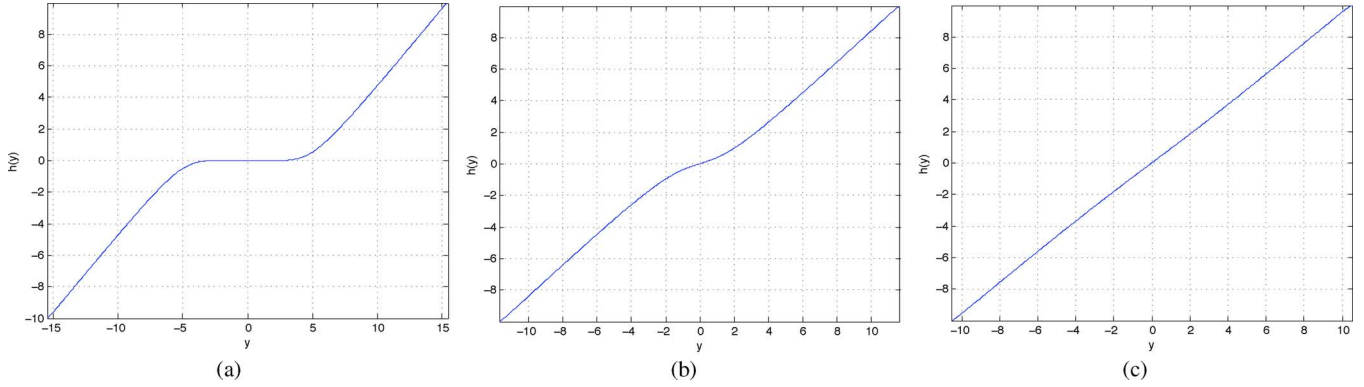
Fig. 2. Optimal estimator at various SNR values when $X \sim \mathcal{N}(0, 1)$ and Z is distributed uniformly on the interval $[-a, a]$. The SNR is varied by changing a. Observe that the optimal estimator converges to linear as SNR increases. (a) $\mathrm{SNR} = 0.1$. (b) $\mathrm{SNR} = 1$. (c) $\mathrm{SNR} = 10$.
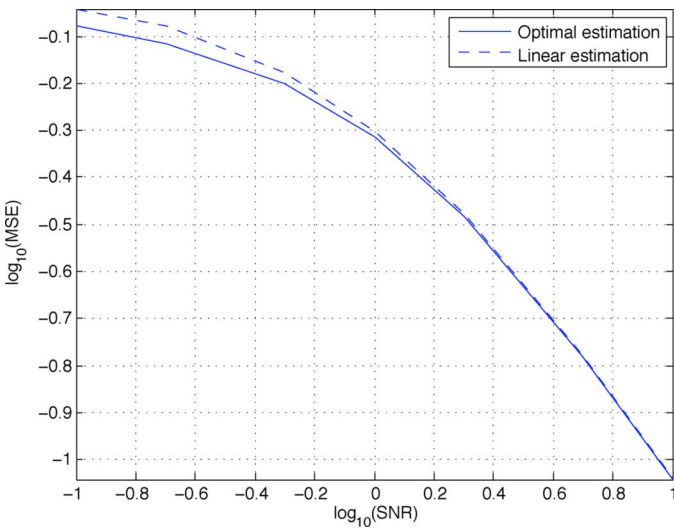

Fig. 3. Variation of estimation error with the channel SNR when $X \sim \mathcal{N}(0, 1)$ and Z is distributed uniformly on the interval $[-a, a]$. We observe that the error is significant at $\gamma = 0.1$ and vanishes at high SNRs.

V. EXTENSION TO VECTOR SPACES

Extension of the conditions to the vector case is nontrivial due to the dependences across components of the source and noise. In this section, for simplicity, we restrict ourselves to the MSE distortion measure. We first give the formal definition of the problem:

We consider the problem of estimating the vector source $X \in \mathbb{R}^m$ given the observation $Y = X + Z$, where X and $Z \in \mathbb{R}^m$ are independent, as shown in Fig. 1. Without loss of generality, we assume that X and Z are zero mean random variables with m-fold distributions $f_X(\cdot)$ and $f_Z(\cdot)$. Their respective characteristic functions are denoted $F_X(\omega)$ and $F_Z(\omega)$. $R_X = \mathbb{E}\{XX^T\}$, $R_Z = \mathbb{E}\{ZZ^T\}$ are the covariance matrices of X and Z, respectively. Let us represent $R_X R_Z^{-1}$ as

$$R_X R_Z^{-1} = U^{-1} \Lambda U. \tag{27}$$

where Λ is a diagonal matrix with elements $\lambda_1, \ldots, \lambda_m$.

We are looking for the conditions on $F_X(\omega)$ and $F_Z(\omega)$ such that $h(Y) = KY$ with $K = R_X(R_X + R_Z)^{-1}$ minimizes the estimation error $\mathbb{E}\{\|X - h(Y)\|_2^2\}$.

By following a similar approach (details are in Appendix D) to the scalar case, we obtain the necessary and sufficient condition of optimality

$$U \nabla \log F_X(\omega) = \Lambda U \nabla \log F_Z(\omega). \tag{28}$$

We will make use of the following auxiliary lemma from matrix analysis.

Lemma 2: Given a function $f : \mathbb{R}^n \to \mathbb{R}$, matrix $A \in \mathbb{R}^{n \times m}$ and vector $x \in \mathbb{R}^m$

$$\nabla_x f(Ax) = A^T \nabla f(Ax). \tag{29}$$

Proof: See Appendix E. ∎

Next, we state the main theorem in the vector setting.

Theorem 6: Let the characteristic functions of the transformed source and noise (UX and UZ) be $F_{UX}(\omega)$ and $F_{UZ}(\omega)$. The necessary and sufficient condition for linearity of optimal estimation is

$$\frac{\partial \log F_{UX}(\omega)}{\partial \omega_i} = \lambda_i \frac{\partial \log F_{UZ}(\omega)}{\partial \omega_i}, \ 1 \le i \le m. \tag{30}$$

Proof: Let us define $\vec{\omega} = (U^{-T})\omega$; hence, $\omega = U^T \vec{\omega}$. Plugging this in (28), we have

$$U \nabla_{U^T \vec{\omega}} \log F_X(U^T \vec{\omega}) = \Lambda U \nabla_{U^T \vec{\omega}} \log F_Z(U^T \vec{\omega}). \tag{31}$$

Using Lemma 2, we can rewrite (31) as

$$\nabla_{\vec{\omega}} \log F_X(U^T \vec{\omega}) = \Lambda \nabla_{\vec{\omega}} \log F_Z(U^T \vec{\omega}). \tag{32}$$

Note that the characteristic functions of the source and noise after transformation can be written in terms of the known characteristic functions $F_X(\omega)$ and $F_Z(\omega)$, specifically $F_{UX}(\omega) = F_X(U^T \omega)$ and $F_{UZ}(\omega) = F_Z(U^T \omega)$. Plugging these expressions in (32), we have

$$\nabla_{\vec{\omega}} \log F_{UX}(\vec{\omega}) = \Lambda \nabla_{\vec{\omega}} \log F_{UZ}(\vec{\omega}). \tag{33}$$

Using the fact that Λ is diagonal, we convert (33) to the set of m scalar differential equations of (30). ∎

Further insight into the earlier necessary and sufficient condition is provided via the following corollaries.

Corollary 6: Let $F_{[UX]_i}(\omega)$ and $F_{[UZ]_i}(\omega)$ be the marginal characteristic functions of the transform coefficients $[UX]_i$ and $[UZ]_i$, respectively. A necessary condition for linearity of optimal estimation is

$$F_{[UX]_i}(\omega) = F_{[UZ]_i}^{\lambda_i}(\omega), \ 1 \leq i \leq m. \qquad (34)$$

Proof: The marginal characteristic functions of $[UX]_i$ and $[UZ]_i$ are obtained by setting $\omega_k = 0$, $\forall k \neq i$ in $F_{UX}(\omega)$ and $F_{UZ}(\omega)$, respectively. By setting $\omega_k = 0$, $\forall k \neq i$ in both sides of (30), we have

$$\frac{\partial \log F_{[UX]_i}(\omega)}{\partial \omega} = \lambda_i \frac{\partial \log F_{[UZ]_i}(\omega)}{\partial \omega}, \ 1 \leq i \leq m. \quad (35)$$

The solution to this differential equation is given by

$$\log F_{[UX]_i}(\omega) = \lambda_i \log F_{[UZ]_i}(\omega) + C \qquad (36)$$

where C is a constant. Imposing $F_{[UZ]_i}(0) = F_{[UX]_i}(0) = 1$, we obtain $C = 0$, which implies

$$F_{[UX]_i}(\omega) = F_{[UZ]_i}^{\lambda_i}(\omega), \ 1 \leq i \leq m. \qquad (37)$$

■

Corollary 7: A necessary condition for linearity of optimal estimation is that one of the following holds for every pair i, j, $1 \leq i, j \leq m$:
1) $\lambda_i = \lambda_j$
2) $[UX]_i$ is independent of $[UX]_j$ and $[UZ]_i$ is independent of $[UZ]_j$.

Proof: Let us rewrite (30) explicitly for the ith and jth coefficients

$$\frac{\partial \log F_{UX}(\omega)}{\partial \omega_i} = \lambda_i \frac{\partial \log F_{UZ}(\omega)}{\partial \omega_i} \qquad (38)$$

$$\frac{\partial \log F_{UX}(\omega)}{\partial \omega_j} = \lambda_j \frac{\partial \log F_{UZ}(\omega)}{\partial \omega_j}. \qquad (39)$$

The partial derivative of both sides of (38) with respect to ω_j and both sides of (39) with respect to ω_i is obtained as follows:

$$\frac{\partial^2 \log F_{UX}(\omega)}{\partial \omega_i \partial \omega_j} = \lambda_i \frac{\partial^2 \log F_{UZ}(\omega)}{\partial \omega_i \partial \omega_j} \qquad (40)$$

$$\frac{\partial^2 \log F_{UX}(\omega)}{\partial \omega_i \partial \omega_j} = \lambda_j \frac{\partial^2 \log F_{UZ}(\omega)}{\partial \omega_i \partial \omega_j}. \qquad (41)$$

There are only two ways to simultaneously satisfy (40) and (41): 1) $\lambda_i = \lambda_j$ 2) the second-order derivatives vanish, i.e.

$$\frac{\partial^2 \log F_{UX}(\omega)}{\partial \omega_i \partial \omega_j} = 0 \qquad (42)$$

$$\frac{\partial^2 \log F_{UZ}(\omega)}{\partial \omega_i \partial \omega_j} = 0. \qquad (43)$$

Let us focus on X i.e., (42), derivation for Z follows similarly. $F_{[UX]_{ij}}(\omega_i, \omega_j)$, i.e., the marginal characteristic function of the pair $([UX]_i, [UX]_j)$ is obtained by setting $\omega_k = 0$, $\forall k \neq i, j$. Then, (42) implies

$$\frac{\partial^2 \log F_{[UX]_{ij}}(\omega_i, \omega_j)}{\partial \omega_i \partial \omega_j} = 0 \qquad (44)$$

which means

$$\log F_{[UX]_{ij}}(\omega_i, \omega_j) = A(\omega_i) + B(\omega_j) \qquad (45)$$

for some functions A and B, i.e., $\log F_{[UX]_{ij}}(\omega_i, \omega_j)$ is additively separable in terms of ω_i and ω_j. This implies

$$F_{[UX]_{ij}}(\omega_i, \omega_j) = C(\omega_i)D(\omega_j) \qquad (46)$$

for some functions C and D. But (46) implies independence of the ith and jth transform coefficients of source X. The independence of the ith and jth transform coefficients of the noise Z follows from similar arguments. ■

Corollary 8: If the necessary condition of Corollary 6 is satisfied, then a sufficient condition for linearity of optimal estimation is that U generates independent coefficients for both X and Z.

Proof: Independence of the transform coefficients implies that the joint characteristic function is the product of the marginals

$$F_{UX}(\omega) = \prod_{i=1}^{m} F_{[UX]_i}(w_i), F_{UZ}(\omega) = \prod_{i=1}^{m} F_{[UZ]_i}(w_i). \quad (47)$$

Plugging (47) into the necessary and sufficient condition (30) of Theorem 6, it is straightforward to show that (34) the necessary condition of Corollary 6 is now both necessary and sufficient. ■

While the condition in Corollary 8 involves independence of transform coefficients, the weaker property of uncorrelatedness is already guaranteed by transform U. The matrix U diagonalizes both R_X and R_Z. We formalize this in the following lemma.

Lemma 3: Transform U decorrelates both source and noise: both UR_XU^T and UR_ZU^T are diagonal matrices.

Proof: Since both R_X and R_Z are, by definition, positive-definite matrices, there exists a matrix S that simultaneously diagonalizes R_X and whitens R_Z, i.e., $SR_XS^T = \Lambda_X$ and $SR_ZS^T = I$ where Λ_X is diagonal and I is the identity matrix [21]. Hence, R_X and R_Z can be expressed as follows:

$$R_X = S^{-1}\Lambda_X S^{-T}, R_Z = S^{-1}S^{-T}. \qquad (48)$$

Plugging (48) into (27), we obtain $U = \Lambda_U S$, where Λ_U is diagonal. Substituting U in UR_XU^T and UR_ZU^T, we obtain

$$UR_XU^T = \Lambda_U \Lambda_X \Lambda_U^T, UR_ZU^T = \Lambda_U \Lambda_U^T. \qquad (49)$$

The product of diagonal matrices is also diagonal. ■

As an example where the optimal estimator is known to be linear, consider the multivariate Gaussian case. Note that the Gaussian source–channel pair satisfies the scalar matching condition for any SNR, i.e., (37). As any linear transform preserves joint Gaussianity in the transform domain, U generates jointly Gaussian and uncorrelated coefficients which are, therefore, independent, satisfying the conditions of Corollary 8.

Another, perhaps surprising, example where the optimal estimator is linear involves identically distributed source X and noise Z. In this case, the linear estimator is optimal irrespective of the distribution of source and noise. It is straightforward to

show that the necessary and sufficient conditions of Theorem 6 are satisfied if $F_X(\omega) = F_Z(\omega)$.

Example 2: Let us consider a numerical example that highlights the differences in conditions derived for vectors from the scalars. Consider a setting where a 2-D random variable Z' has independent components, both of which are uniformly distributed over $[-a, a]$, i.e., $Z' = [Z'_1, Z'_2]$ and $Z'_1 \sim Z'_2 \sim U[-a, a]$. Also, let X' have two independent identically distributed components $X' = [X'_1, X'_2]$ where X'_1 and X'_2 are distributed according to a density given by the convolution of the uniform density with itself, i.e., $X'_1 \sim X'_2 \sim (U[-a, a] * U[-a, a])$. Since X' and Z' satisfy the sufficient conditions in Corollary 8, the optimal estimator is linear for the source–channel pair (X', Z').

Let us next consider the source–channel pair (X, Z) to be $X = Q_X X'$ and $Z = Q_Z Z'$ where Q_X and Q_Z are 2×2 orthogonal matrices ($Q_X Q_X^T = Q_Z Q_Z^T = I$). This introduces dependences among the components of X and Z. We already saw that for $Q_X = Q_Z = I$, the optimal estimator is linear. Also, from standard linear estimation principles [22], it follows that the minimum estimation error achievable by linear estimators does not depend on Q_X and Q_Z, i.e., linear estimation error is a constant with respect to Q_X and Q_Z. The question we are interested in is—can the linear estimator be optimal for any other pair (Q_X, Q_Z)? Corollary 8 sheds light on this question. First, we consider the case where $Q_X = \pm Q_Z$. Observe that, any orthogonal matrix U satisfies condition (27). Hence, we can set $U = Q_X^{-1} = \pm Q_Z^{-1}$ leading to $UX = X'$ and $UZ = Z'$. This implies that UX and UZ satisfy conditions in Corollary 8, which are sufficient to prove linearity of optimal estimators. Hence, for the source–channel pair (X, Z), optimal estimators are always linear if $Q_X = \pm Q_Z$.

Finally, we consider the case where $Q_X \neq \pm Q_Z$. In general, any orthogonal matrix can be written in terms of another orthogonal matrix as

$$Q_X = G(\theta) Q_Z \qquad (50)$$

where $G(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$ (also known as Givens rotation [21]). For a constant Q_Z, we change Q_X by varying θ and observe the behavior of the difference between the MSEs obtained by the optimal and the linear estimators. As a performance metric, we consider the normalized difference of estimation errors, i.e., (MSE of linear estimation − MSE of optimal estimation)/ MSE of optimal estimation. The variation of the normalized difference as a function of θ is plotted in Fig. 4. Observe that, at $\theta = 0$ and π, the optimal estimator is linear as expected from Corollary 8. It is not hard to show using symmetry of X' and Z' that the conditions of Corollary 8 are also satisfied for $\theta = \pi/2$ (and $3\pi/2$). A perhaps interesting observation is that the deviation of optimal estimator from linearity grows monotonically in θ in the range $\theta \in (0, \pi/4)$.

An important observation is that the necessary and sufficient condition for scalars (14) is also a necessary condition for vectors (34) in the transform domain. Due to this fact, it is straightforward to extend the existence and uniqueness results and implications of the scalar matching conditions to the vector spaces. These trivial extensions are omitted here for conciseness.
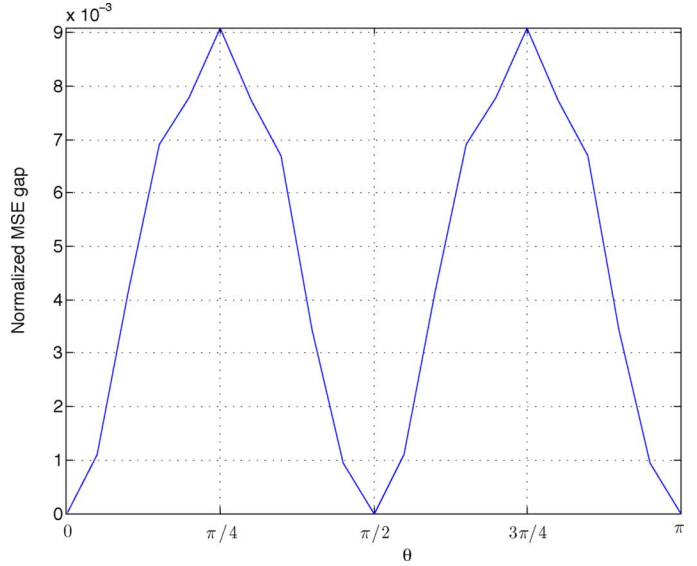

Fig. 4. Normalized difference between optimal and linear estimation versus the Givens rotation parameter θ, for the source–channel pair (X, Z).

VI. CONCLUSION

In this paper, we derived conditions under which the L_p optimal estimator is linear. We identified the conditions for the existence and uniqueness of a source distribution that matches the noise in a way that ensures linearity of the optimal estimator, for the special case of $p = 2$. One trivial example of this type of matching occurs for Gaussian source and Gaussian noise at all SNR levels. Another instance of matching happens when the source and noise are identically distributed. We also showed that the Gaussian source–channel pair is unique in that it is the only pair for which the optimal estimator is linear at more than one SNR values. Moreover, we showed the asymptotic linearity of MSE optimal estimators at low SNR if the channel is Gaussian, regardless of the source, and vice versa, at high SNR if the source is Gaussian regardless of the channel. We also studied the extension to vector spaces where additional conditions are derived beyond those inherited from the scalar case, which concern interactions across components.

APPENDIX A
PROOF OF LEMMA 1

Proof: First, we show the sufficiency of the necessary conditions for L_p norm. Note that $\Phi(x) = |x|^p$ is convex for $p \geq 2$, i.e., $\frac{d^2 |x|^p}{dx^2} \geq 0, \ \forall x \in R - \{0\}$. We need to show $\frac{\partial^2}{\partial^2 \epsilon} J[h(y) + \epsilon \eta(y)]\Big|_{\epsilon=0} \geq 0$, for any $\eta(y)$ variation function

$$\frac{\partial^2}{\partial^2 \epsilon} J[h(y) + \epsilon \eta(y)]\Big|_{\epsilon=0} =$$
$$\int \int \eta^2(y) \Phi''(x - h(y)) f_X(x) f_Z(y - x) dx dy. \quad (51)$$

All factors in the integral are nonnegative, and hence, $\frac{\partial^2}{\partial^2 \epsilon} J[h(y) + \epsilon \eta(y)]\Big|_{\epsilon=0} \geq 0$, for any $\eta(y)$.

Next, we show the uniqueness (in probabilistic sense) of the optimal estimator for even natural p. Assume $h_1(Y)$ and $h_2(Y)$ both satisfy (6) while $\mathbb{P}[h_1(Y) \neq h_2(Y)] > 0$, i.e., over a set of positive measure $h_1(Y) \neq h_2(Y)$. Then, the following holds for any $\eta(Y)$:

$$\mathbb{E}\left\{\{[X - h_2(Y)]^{p-1} - [X - h_1(Y)]^{p-1}\}\eta(Y)\right\} = 0. \quad (52)$$

Note that

$$[X - h_2(Y)]^{p-1} - [X - h_1(Y)]^{p-1} = (h_1(Y) - h_2(Y))\beta(X, Y) \quad (53)$$

where

$$\beta(X, Y) = \sum_{m=0}^{p-2} [X - h_1(Y)]^{p-1-m}[X - h_2(Y)]^m. \quad (54)$$

Proposition: $h_1(Y) \neq h_2(Y)$ implies $\beta(X, Y) > 0 \; \forall X, Y \in \mathbb{R}$.

To see this, we note that (53) is a simple factorization of the form

$$A^{p-1} - B^{p-1} = (A - B)P(A, B) \quad (55)$$

where $P(A, B)$ is a polynomial. Now if $A \neq B$, then the sign of left-hand side equals to the sign of $A - B$. Hence, $P(A, B) > 0$.

Next, plugging $\eta(Y) = h_1(Y) - h_2(Y)$ in (52), we obtain

$$\mathbb{E}\left\{[h_1(Y) - h_2(Y)]^2 \beta(X, Y)\right\} = 0. \quad (56)$$

Since $h_1(Y) \neq h_2(Y)$ implies $\beta(X, Y) > 0 \; \forall X, Y \in \mathbb{R}$, then (56) requires $h_1(Y) = h_2(Y)$ almost everywhere, contradicting the hypothesis $\mathbb{P}[h_1(Y) \neq h_2(Y)] > 0$. ∎

APPENDIX B
PROOF OF THEOREM 1

The necessary and sufficient condition (6) can be rewritten as

$$\int \left\{\int (x - ky)^{p-1} f_X(x) f_Z(y - x)dx\right\}\eta(y)dy = 0 \quad (57)$$

for all admissible perturbation functions $\eta(y)$. This equality is achieved for all $\eta(y)$ if and only if the expression in braces vanishes almost everywhere. Hence, (6) is satisfied if and only if

$$\Psi(y) \triangleq \int (x - ky)^{p-1} f_X(x) f_Z(y - x)dx = 0, a.e. \quad (58)$$

Applying the binomial expansion to the first factor

$$(x - ky)^{p-1} = \sum_{m=0}^{p-1} \binom{p-1}{m} (-ky)^m x^{p-m-1} \quad (59)$$

and rearranging terms, we get

$$\sum_{m=0}^{p-1} \binom{p-1}{m} (-ky)^m \int x^{p-1-m} f_X(x) f_Z(y - x)dx = 0. \quad (60)$$

Let $*$ denote the convolution operator, and rewrite (60) as

$$\sum_{m=0}^{p-1} \binom{p-1}{m} (-ky)^m \left[y^{p-1-m} f_X(y) * f_Z(y)\right] = 0. \quad (61)$$

Taking the Fourier transform[6]

$$\sum_{m=0}^{p-1} \binom{p-1}{m} (-k)^m \frac{d^m}{d\omega^m}\left[\frac{d^{p-1-m}(F_X(\omega))}{d\omega^{p-1-m}} F_Z(\omega)\right] = 0 \quad (62)$$

differentiating in parts

$$\sum_{m=0}^{p-1} \binom{p-1}{m} (-k)^m \sum_{l=0}^{m} \binom{m}{l} \frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}} \frac{d^l F_Z(\omega)}{d\omega^l} = 0 \quad (63)$$

interchanging summations

$$\sum_{l=0}^{p-1} \frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}} \frac{d^l F_Z(\omega)}{d\omega^l} \sum_{m=l}^{p-1} \binom{p-1}{m}(-k)^m \binom{m}{l} = 0 \quad (64)$$

applying some combinatoric algebra

$$\sum_{l=0}^{p-1} \binom{p-1}{l} \frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}} \frac{d^l F_Z(\omega)}{d\omega^l}$$
$$\sum_{m=l}^{p-1} \frac{(p-1-l)!}{(m-l)!(p-1-m)!}(-k)^m = 0 \quad (65)$$

and substituting $t = m - l$, we get

$$\sum_{l=0}^{p-1} \binom{p-1}{l} \frac{d^{p-1-l}F_X(\omega)}{d\omega^{p-1-l}} \frac{d^l F_Z(\omega)}{d\omega^l}$$
$$\sum_{t=0}^{p-1-l} \binom{p-1-l}{t}(-k)^{(t+l)} = 0. \quad (66)$$

Finally, noting that

$$(1 - k)^{p-1-l} = \sum_{t=0}^{p-1-l} \binom{p-1-l}{t}(-k)^t \quad (67)$$

we obtain that (10) is a necessary and sufficient condition.

We note that all steps of the derivation were obtained as " if and only if " statements; hence, the converse is automatically proved.

[6]Note that the Fourier transforms exist due to the finite moments assumption stated in Section II-A.

APPENDIX C
FORMAL PROOF OF THEOREM 5

Let $h(y) = k[y + \xi(y)]$ be the polynomial expansion of the optimal estimator where $\xi(y)$ consists of terms with order only two or higher. Let us rewrite the optimal estimator

$$h(y) = k[y + \xi(y)] = \frac{\int x f_X(x) f_Z(y-x) \, dx}{\int f_X(x) f_Z(y-x) \, dx} \qquad (68)$$

or

$$k[y+\xi(y)] \int f_X(x) f_Z(y-x) \, dx = \int x f_X(x) f_Z(y-x) \, dx. \qquad (69)$$

Expressing the integrals as convolutions, we have

$$k[y + \xi(y)] [f_X(y) * f_Z(y)] = [y f_X(y)] * f_Z(y). \qquad (70)$$

Taking the Fourier transform of both sides, we obtain

$$jk \frac{d[F_X(\omega)F_Z(\omega)]}{dw} + k[F_X(\omega)F_Z(\omega)] * \Xi(\omega)$$
$$= jF_Z(\omega)\frac{dF_X(\omega)}{dw} \qquad (71)$$

where $\Xi(\omega)$ denotes the Fourier transform of $\xi(\cdot)$. Plugging $k = \frac{\gamma}{1+\gamma}$ and dividing both sides by $F_X(\omega)F_Z(\omega)$ we have

$$\frac{1}{F_X(\omega)} \frac{dF_X(\omega)}{d\omega} = \frac{\gamma}{1+\gamma}\zeta(\omega) + \gamma \frac{1}{F_Z(\omega)}\frac{dF_Z(\omega)}{d\omega} \qquad (72)$$

or more compactly

$$\frac{d}{d\omega} \log F_X(\omega) = \frac{\gamma}{1+\gamma}\zeta(\omega) + \frac{d}{d\omega} \log F_Z^\gamma(\omega) \qquad (73)$$

where $\zeta(\omega) \triangleq \frac{\gamma}{1+\gamma}\frac{[F_X(\omega)F_Z(\omega)]*\Xi(\omega)}{j[F_X(\omega)F_Z(\omega)]}$.

Now, consider the setting where the source is Gaussian and $\gamma \to \infty$. By applying the central limit theorem, we have $F_Z^\gamma(\omega) \to F_X(\omega)$ pointwise as $\gamma \to \infty$. Hence, $\zeta(\omega) \to 0$ pointwise for all $\omega \in \mathbb{R}$. But this implies $\Xi(\omega) \to 0$ (pointwise) and hence in the limit $\gamma \to \infty$, $\xi(y) = 0$ almost everywhere with respect to the density of y. Also, it follows from the same arguments that when the noise is Gaussian and $\gamma \to 0$, $\xi(y) = 0$ a.e.

APPENDIX D
DERIVATION-VECTOR CASE

Let us rewrite the MSE optimal estimator for the vector case

$$\boldsymbol{h(y)} = \frac{\int \boldsymbol{x} f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - \boldsymbol{x}) \, d\boldsymbol{x}}{\int f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - \boldsymbol{x}) \, d\boldsymbol{x}}. \qquad (74)$$

Plugging $\boldsymbol{h(y)} = K\boldsymbol{y}$ in (74) we obtain

$$K\boldsymbol{y} \int f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - \boldsymbol{x}) \, d\boldsymbol{x} = \int \boldsymbol{x} f_X(\boldsymbol{x}) f_Z(\boldsymbol{y} - \boldsymbol{x}) \, d\boldsymbol{x}. \qquad (75)$$

Expressing the integrals as m-fold convolutions, we get

$$K\boldsymbol{y} [f_X(\boldsymbol{y}) * f_Z(\boldsymbol{y})] = [\boldsymbol{y} f_X(\boldsymbol{y})] * f_Z(\boldsymbol{y}). \qquad (76)$$

Taking the Fourier transform of both sides

$$jK\nabla [F_X(\omega)F_Z(\omega)] = jF_Z(\omega)\nabla F_X(\omega) \qquad (77)$$

and rearranging terms, we get

$$(I - K)\frac{1}{F_X(\omega)}\nabla F_X(\omega) = K\frac{1}{F_Z(\omega)}\nabla F_Z(\omega). \qquad (78)$$

Using $\nabla \log F_X(\omega) = \frac{1}{F_X(\omega)}\nabla F_X(\omega)$

$$\nabla \log F_X(\omega) = (I - K)^{-1}K\nabla \log F_Z(\omega). \qquad (79)$$

Note that (see, e.g., [22])

$$K = R_X(R_X + R_Z)^{-1} \qquad (80)$$

hence we have

$$(I - K) = (R_X + R_Z)(R_X + R_Z)^{-1} - R_X(R_X + R_Z)^{-1}$$
$$= R_Z(R_X + R_Z)^{-1} \qquad (81)$$

and

$$(I - K)^{-1}K = [R_Z(R_X + R_Z)^{-1}]^{-1}R_X(R_X + R_Z)^{-1}$$
$$= [R_Z(R_X + R_Z)^{-1}]^{-1}[R_X + R_Z - R_Z](R_X + R_Z)^{-1}$$
$$= [R_Z(R_X + R_Z)^{-1}]^{-1} - I$$
$$= [(R_X + R_Z)R_Z^{-1}] - I$$
$$= R_X R_Z^{-1} + I - I$$
$$= R_X R_Z^{-1}. \qquad (82)$$

Plugging (82) into (79) we obtain

$$\nabla \log F_X(\omega) = R_X R_Z^{-1}\nabla \log F_Z(\omega). \qquad (83)$$

Using the eigendecomposition of $R_X R_Z^{-1} = U^{-1}\Lambda U$ where Λ is diagonal with eigenvalues $\lambda_1, \ldots, \lambda_n$, we obtain

$$U\nabla \log F_X(\omega) = \Lambda U\nabla \log F_Z(\omega). \qquad (84)$$

Similar to the scalar case, we can show the converse by retracing the steps in the derivation of the necessity. Note that none of these steps, (74)–(84), introduce any loss of generality; hence, retracing back from (84) to (74), we show that if (84) is satisfied, the optimal estimator is linear.

APPENDIX E
PROOF OF LEMMA 2

By the chain rule, we have

$$\frac{\partial f(A\boldsymbol{x})}{\partial x_i} = \sum_{k=1}^n \frac{\partial f(A\boldsymbol{x})}{\partial [A\boldsymbol{x}]_k} \frac{\partial [A\boldsymbol{x}]_k}{\partial [\boldsymbol{x}]_i} \qquad (85)$$

$$= \sum_{k=1}^n \frac{\partial f(A\boldsymbol{x})}{\partial [A\boldsymbol{x}]_k} \frac{\partial ([A]_k^T \boldsymbol{x})}{\partial [\boldsymbol{x}]_i} \qquad (86)$$

$$= \sum_{k=1}^n \frac{\partial f(A\boldsymbol{x})}{\partial [A\boldsymbol{x}]_k} [A]_{ki} \qquad (87)$$

$$= \sum_{k=1}^{n} \partial_k f(A\boldsymbol{x})[A]_{ki} \qquad (88)$$

$$= [A]_i^T \nabla f(A\boldsymbol{x}). \qquad (89)$$

It follows from (89) that $\nabla_x f(A\boldsymbol{x}) = A^T \nabla f(A\boldsymbol{x})$.

ACKNOWLEDGMENT

The authors would like to thank the editor and the referees for their comments that helped to clarify the results.

REFERENCES

[1] H. V. Allen, "A theorem concerning the linearity of regression," *Statist. Res. Memoirs*, vol. 2, pp. 60–68, 1938.

[2] C. Rothschild and E. Mourier, "Sur les lois de probabilitéà regression linéaire etécart type lié constant," *Comptes Rendus*, vol. 225, pp. 1117–1119, 1947.

[3] M. M. Rao and R. J. Swift, *Probability Theory With Applications*. New York: Springer, 2005.

[4] S. G. Ghurye and I. Olkin, "A characterization of the multivariate normal distribution," *Ann. Math. Statist.*, pp. 533–541, 1962, vol. 33.

[5] P. Billingsley, *Probability and Measure*. New York: Wiley, 2008.

[6] C. R. Rao, "Note on a problem of Ragnar Frisch," *Econometrica, J. Econometric Soc.*, vol. 15, no. 3, pp. 245–249, 1947.

[7] E. Akyol, K. Viswanatha, and K. Rose, "On conditions for linearity of optimal estimation," in *Proc. IEEE Inf. Theory Workshop*, 2010, pp. 1–5.

[8] E. Akyol, K. Viswanatha, and K. Rose, "On multidimensional optimal estimators: Linearity conditions," in *Proc. IEEE Statist. Signal Process. Workshop*, 2011, pp. 741–744.

[9] R. G. Laha, "On a characterization of the stable law with finite expectation," *Ann. Math. Statist.*, vol. 27, no. 1, pp. 187–195, 1956.

[10] A. Balakrishnan, "On a characterization of processes for which optimal mean-square systems are of specified form," *IEEE Trans. Inf. Theory*, vol. 6, no. 4, pp. 490–500, Sep. 1960.

[11] D. G. Luenberger, *Optimization by Vector Space Methods*. New York: Wiley, 1969.

[12] S. Sherman, "Non-mean-square error criteria," *IEEE Trans. Inf. Theory*, vol. 4, no. 3, pp. 125–126, Sep. 1958.

[13] E. Lukacs, *Characteristics Functions*. London, U.K.: Griffin, 1960.

[14] F. W. Steutel and K. Van Harn, *Infinite Divisibility of Probability Distributions on the Real Line*. Boca Raton, FL: CRC Press, 2003.

[15] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*. Providence, RI: American Mathematical Society, 1943.

[16] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. New York: Springer, 1992.

[17] R. M. Gray and T. G. Stockham, Jr, "Dithered quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, May 1993.

[18] C. R. Rao, "On some characterizations of the normal law," *Indian J. Statist., Ser. A*, vol. 29, no. 1, pp. 1–14, 1967.

[19] C. D. Hardin, "On the linearity of regression," *Probabil. Theory Related Fields*, vol. 61, no. 3, pp. 293–302, 1982.

[20] R. M. Dudley, *Real Analysis and Probability*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[21] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[22] T. Kailath, A. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 2000.

Emrah Akyol (S'03) received the B.Sc. degree in 2003 from Bilkent University (Turkey), the M.Sc. degree in 2005 from Koc University (Turkey), and the Ph.D. degree in 2011 in electrical and computer engineering from the University of California at Santa Barbara. From 2006 to 2007, he held positions at Hewlett-Packard Laboratories and NTT Docomo Laboratories, both in Palo Alto, CA, where he worked on topics in video compression.

Currently, Dr. Akyol is a postdoctoral researcher in the Department of Electrical and Computer Engineering, University of California at Santa Barbara. His research focuses on source and source-channel coding, energy efficient communications, multimedia compression and networking, and the connections between estimation theory and information theory.

Kumar B. Viswanatha (S'08) received his B.Tech. in electrical engineering in 2008 from the Indian Institute of Technology – Madras (IIT – Madras), Chennai, India and his M.S. in electrical and computer engineering in 2009 from University of California at Santa Barbara (UCSB), USA. He is currently pursuing his Ph.D. in electrical and computer engineering at UCSB. He was an intern associate in the equity volatility desk at Goldman Sachs Co., New York. His research interests include multi-user information theory, joint compression and routing for networks and distributed compression for large scale sensor networks.

Kenneth Rose (S'85–M'91–SM'01–F'03) received the Ph.D. degree in 1991 from the California Institute of Technology, Pasadena.

He then joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, where he is currently a Professor. His main research activities are in the areas of information theory and signal processing, and include rate-distortion theory, source and source-channel coding, audio and video coding and networking, pattern recognition, and non-convex optimization. He is interested in the relations between information theory, estimation theory, and statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines.

Dr. Rose was co-recipient of the 1990 William R. Bennett Prize Paper Award of the IEEE Communications Society, as well as the 2004 and 2007 IEEE Signal Processing Society Best Paper Awards.