On the Structure of the Least Favorable Prior Distributions

Alex Dytso*, Ronit Bustin**, H. Vincent Poor*, and Shlomo Shamai (Shitz)**

***Abstract*—This paper studies optimization of the minimum mean square error (MMSE) in order to characterize the structure of the least favorable prior distributions. In the first part, the paper characterizes the local behavior of the MMSE in terms of the input distribution and finds the directional derivative of the MMSE at the distribution $P_{\mathbf{X}}$ in the direction of the distribution $Q_{\mathbf{X}}$.**

In the second part of the paper, the directional derivative together with the theory of convex optimization is used to characterize the structure of least favorable distributions. In particular, under mild regularity conditions, it is shown that the support of the least favorable distributions must necessarily be very small and is contained in a nowhere dense set of Lebesgue measure zero. The results of this paper produce both sufficient and necessary conditions for optimality, do not rely on Gaussian statistics assumptions, and are not sensitive to the dimensionality of random vectors. The results are evaluated for the univariate and multivariate random Gaussian cases, and the Poisson case. Finally, as one of the applications, it is shown how the results can be used to characterize the capacity of Gaussian MIMO channels with an amplitude constraint.

I. Introduction

The *minimum mean square error* (MMSE) in estimating an input random vector $\mathbf{X} \in \mathbb{R}^n$ from a noisy observation/output $\mathbf{Y} \in \mathbb{R}^k$ is defined as

$$\mathrm{mmse}(\mathbf{X}|\mathbf{Y}) \triangleq \inf_{f(\cdot):f \text{ is measurable}} \mathbb{E}\left[\|\mathbf{X}-f(\mathbf{Y})\|^2\right]. \quad (1)$$

In this paper we study the problem of maximizing the MMSE in (1) over the set of input distributions on $\mathbf{X}$ for a fixed transition distribution $P_{\mathbf{Y}|\mathbf{X}}$. Specifically, we will work with the following two sets: 1) the set of distributions with a compact support; and 2) the set of distributions with finite generalized moments (e.g., second moment, third absolute moment, logarithmic moments, etc.). The distributions that achieve the worst-case MMSE (i.e., maximize the MMSE) are called *least favorable prior distributions*.

The problem of finding least favorable prior distributions is interesting from both *estimation theoretic* and *information theoretic* points of view. Firstly, in estimation theory, maximization of the MMSE over a set of distributions with compact support is directly relate to the problem of characterizing a minimax estimator [1]. Specifically, a conditional expectation (optimal Bayes estimator) evaluated with a least favorable prior distribution is also a *minimax estimator*.

Secondly, in information theory, in view of the I-MMSE relationship [2] that connects the MMSE and the mutual information for the case of additive Gaussian noise, the least favorable distributions are often also capacity achieving distributions (i.e., maximize mutual information). For example, in [3] such an approach was used to characterize the capacity achieving distribution of a Gaussian noise channel with a small (but nonvanishing) input amplitude constraint.

Unlike previous works, the approach taken in this work is based on the theory of convex optimization and allows us to produce systematic and general results. For instance, our approach produces both sufficient and necessary conditions for optimality, does not rely on the assumption of Gaussian statistics, and is not sensitive to the dimensionality of random vectors $\mathbf{X}$ and $\mathbf{Y}$. Our approach also parallels the variational approach, used in information theory [4], [5], for finding capacity achieving distributions.

A. Past Work

The theory of finding least favorable prior distributions has received considerable attention under the assumption of univariate and/or Gaussian statistics. For the univariate case under some mild condition, Ghosh in [6] has shown that, with the support constraint on the input, the least favorable priors are discrete with finitely many points. However, as was pointed out in [6] it is not clear how to generalize the argument to the multivariate case. In contrast, the approach taken in this paper is insensitive to the dimensionality.

In [7] for the Gaussian case, capitalizing on the result of Ghosh, the authors demonstrated necessary and sufficient conditions for the optimality of a two point prior distribution. In addition, the authors in [7] also provided a sufficient condition for the optimality of a three point prior. In contrast, the methodology used in this paper produces both sufficient and necessary conditions that can be tested against any N-point prior.

For the multivariate Gaussian case, with a sufficiently small ball constraint, in [1] it has been shown that the least favorable prior is distributed on the boundary of the ball. For a comprehensive overview of the minimax estimation of a bounded mean the interested reader is referred to [8] and references therein.

B. Outline and Paper Contributions

Our contributions are as follows. In Section II we review important properties of the MMSE needed in our analysis. In Section III we characterize the local behavior of the MMSE in terms of the input distribution and find the directional derivative of the MMSE functional at the distribution $P_{\mathbf{X}}$ in the direction of the distribution $Q_{\mathbf{X}}$.

In Section IV we apply the theory of convex optimization to maximize the MMSE. In Section IV-A and Section IV-B we present required mathematical tools such as theorems from convex optimization and theorems of analytic functions. In Section IV-C we look at the case of the compact

*A. Dytso and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA (email: adytso, poor@princeton.edu).

**R. Bustin and S. Shamai (Shitz) are with the Department of Electrical Engineering, Technion-Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: bustin@technion.ac.il, sshlomo@ee.technion.ac.il).

The work of A. Dytso and H. V. Poor was supported by the U. S. National Science Foundation under Grants CCF-1420575, ECCS-1549881 and EECS-1647198. The work of S. Shamai and R. Bustin was supported by the European Union's Horizon 2020 Research and Innovation Programme Grant 694630.

978-1-5386-4780-6/18/$31.00©2018 IEEE

support constraint: Theorem 6 shows that a least favorable input distribution exists for an arbitrary $P_{\mathbf{Y}|\mathbf{X}}$ and derives necessary and sufficient conditions for its optimality; Proposition 1, under some mild conditions, characterizes the structure of the support of least favorable prior distributions and shows that the support must be a *nowhere dense set of Lebesgue measure zero*; Proposition 3 and Proposition 4 look at univariate and multivariate Gaussian noise cases and recover and expand on some known results; Proposition 5 shows how our results can be applied to characterize the capacity of MIMO channels. Surprisingly, Proposition 5 also characterizes the capacity of the MIMO amplitude channel in a regime where the number of antennas approaches infinity; and Proposition 6 considers the Poisson noise case. Section IV-F, looks at least favorable priors under the generalized moment constraints. Section V concludes the paper.

Due to space limitations, some of the proofs are omitted and can be found in an extended version of this paper [9].

C. Notation

Throughout the paper we adopt the following notational conventions: Deterministic scalar quantities are denoted by lowercase letters and deterministic vector quantities are denoted by lowercase bold letters; matrices are denoted by bold uppercase letters; random variables are denoted by uppercase letters and random vectors are denoted by bold uppercase letters; and we denote an n-dimensional ball of radius R centered at $\mathbf{0}$ as $\mathcal{B}_{\mathbf{0}}(R) \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \le R\}$. For a random vector $\mathbf{X}$ with distribution $P_{\mathbf{X}}$ we define the expected value as $\mathbb{E}[\mathbf{X}] = \int \mathbf{x} dP_{\mathbf{X}}(\mathbf{x})$ when we need to emphasize that $\mathbf{X}$ is distributed according to $P_{\mathbf{X}}$ we use the notation $\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}]$. We say that a random vector $\mathbf{Y} \in L^p$ if $\mathbb{E}[\|\mathbf{Y}\|^p] < \infty$; we denote the set of all probability distributions on $\mathsf{S} \subset \mathbb{R}^n$ as $\mathcal{F}_\infty(\mathsf{S})$; and a point $\mathbf{x} \in \mathbb{R}^n$ is said to be a *point of increase* of a distribution $P_{\mathbf{X}}$, if for any open subset $O \subset \mathbb{R}^n$ containing $\mathbf{x}$, $P_{\mathbf{X}}(O) > 0$. We denote the set of points of increase of $P_{\mathbf{X}}$ as $\mathcal{E}(P_{\mathbf{X}}) \subseteq \mathbb{R}^n$. Observe that $P_{\mathbf{X}}(\mathcal{E}(P_{\mathbf{X}})) = 1$. In fact, $\mathcal{E}(P_{\mathbf{X}})$ is the minimal closed subset of $\mathbb{R}^n$ whose probability is 1.

II. The MMSE

In this section we review some important properties of the MMSE.

A. Fundamental Theorems of MMSE Estimation

Theorem 1. (Fundamental Theorems of MMSE Estimation.)

1) (Pythagorean Theorem.) *For any* $f : \mathbb{R}^k \to \mathbb{R}^n$

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right] = \mathbb{E}\left[\|\mathbf{X} - f(\mathbf{Y})\|^2\right] - \mathbb{E}\left[\|f(\mathbf{Y}) - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right]. \quad \text{(2a)}$$

2) (Conditional Expectation is the Optimal Estimator.)

$$\begin{aligned}\mathrm{mmse}(\mathbf{X}|\mathbf{Y}) &= \inf_{f(\cdot): f \textit{ is measurable}} \mathbb{E}\left[\|\mathbf{X} - f(\mathbf{Y})\|^2\right] \\ &= \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right]. \quad \text{(2b)}\end{aligned}$$

B. The MMSE as a Functional

Throughout the paper we will treat the MMSE as an operator (or a functional) on the space of joint distributions $P_{\mathbf{XY}}$. To emphasize that the MMSE is a function of the pair $(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}})$ we use the following notation:

$$\mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \triangleq \mathrm{mmse}(\mathbf{X}|\mathbf{Y}). \quad (3)$$

Continuity of the MMSE will play a key role in our analysis and, therefore, we need the following definitions.

Definition 1. *A function* $f : \mathcal{F} \mapsto \mathbb{R}$ *is said to be upper-semicontinuous (resp. lower-semicont.) at a point* x_0 *if*

$$\limsup_{x \to x_0} f(x) \le f(x_0) \quad \left(\textit{resp.} \ \liminf_{x \to x_0} f(x) \ge f(x_0)\right).$$

A function f *is continuous at* x_0 *if it is both upper and lower semicontinuous at* x_0.

Next, we summarize operator properties of the MMSE.

Theorem 2. (Operator Properties of the MMSE [10].)

1) (Concavity.) $P_{\mathbf{XY}} \mapsto \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}})$ *is a concave functional of* $P_{\mathbf{XY}}$*. Therefore, the MMSE is also concave in* $P_{\mathbf{X}}$ *(resp.* $P_{\mathbf{Y}|\mathbf{X}}$*) if* $P_{\mathbf{Y}|\mathbf{X}}$ *(resp.* $P_{\mathbf{X}}$*) is fixed.*

2) (Upper Semicontinuity.)

- $P_{\mathbf{XY}} \mapsto \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}})$ *is upper semicontinuous over* $\mathcal{M}(\mathsf{S})$ *where* $\mathsf{S} \subseteq \mathbb{R}^n$ *is bounded and*

$$\mathcal{M}(\mathsf{S}) \triangleq \{P_{\mathbf{XY}} : \forall P_{\mathbf{Y}|\mathbf{X}} \textit{ and } P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S})\}.$$

- *Let* $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ *where* $\mathbb{E}[\|\mathbf{N}\|^2] < \infty$*; then* $P_{\mathbf{X}} \mapsto \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}})$ *is upper semicontinuous.*

3) (Continuity.) *Let* $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ *where* $\mathbf{N}$ *has a continuous and bounded density and* $\mathbb{E}[\|\mathbf{N}\|^2] < \infty$*; then* $P_{\mathbf{X}} \mapsto \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}})$ *is continuous.*

III. Local Behavior of the MMSE in Terms of The Input Distribution

Let $P_{\mathbf{X}}$ be the distribution of $\mathbf{X}$. In this section, we study the local behavior of the MMSE as a function of $P_{\mathbf{X}}$.

Definition 2. (The Gâteaux Derivative.) *Let* $\mathcal{F}$ *be a convex topological space. For any two distributions* $P \in \mathcal{F}$ *and* $Q \in \mathcal{F}$ *we define the Gâteaux derivative of a function* $g : \mathcal{F} \to \mathbb{R}$ *at* P *in the direction of* Q *as*

$$\Delta_Q g(P) \triangleq \lim_{\lambda \to 0} \frac{g((1-\lambda)P + \lambda Q) - g(P)}{\lambda}. \quad (4)$$

The Gâteaux derivative is the generalization of a concept of directional derivative and is an important optimization tool. The following theorem finds the Gâteaux derivative of the MMSE with respect to the input distribution.

Theorem 3. (The Gâteaux Derivative of the MMSE.) *For any* $P_{\mathbf{X}}, Q_{\mathbf{X}}$ *and* $P_{\mathbf{Y}|\mathbf{X}}$ *we have that*

$$\begin{aligned}&\Delta_{Q_{\mathbf{X}}} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\ &\quad = \mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) - \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}), \quad \text{(5a)}\end{aligned}$$

where

$$\mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \triangleq \mathbb{E}_{Q_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right]. \quad \text{(5b)}$$

Proof. Let $P_\lambda = (1-\lambda)P_{\mathbf{X}} + \lambda Q_{\mathbf{X}}$. From the definition of the Gâteaux derivative in (4) we have to consider

$$
\begin{aligned}
&\mathrm{mmse}(P_\lambda, P_{\mathbf{Y}|\mathbf{X}}) - \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\
&= \mathbb{E}_{P_\lambda}\left[\|\mathbf{X} - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right] - \mathbb{E}_{P_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\
&\overset{a)}{=} \mathbb{E}_{P_\lambda}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] - \mathbb{E}_{P_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\
&- \mathbb{E}_{P_\lambda}[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2] \\
&\overset{b)}{=} (1-\lambda)\mathbb{E}_{P_{\mathbf{X}}}[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2] \\
&+ \lambda\mathbb{E}_{Q_{\mathbf{X}}}[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2] - \mathbb{E}_{P_{\mathbf{X}}}[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2] \\
&- \mathbb{E}_{P_\lambda}[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2] \\
&= \lambda\mathbb{E}_{Q_{\mathbf{X}}}[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2] - \lambda\mathbb{E}_{P_{\mathbf{X}}}[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2] \\
&- \mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right], \qquad (6)
\end{aligned}
$$

where the steps follow from: a) using the Pythagorean identity in (2a); and b) using the property that the expected value is a linear operator on a set of distributions. Next by dividing (6) by λ and taking $\lambda \to 0$ we have that

$$
\begin{aligned}
&\Delta_{Q_{\mathbf{X}}}\mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\
&= \mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) - \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\
&- \lim_{\lambda\to 0}\frac{\mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right]}{\lambda}. \qquad (7)
\end{aligned}
$$

The proof that the limit in (7) is zero is delegate to [9]. □

IV. Optimization of the MMSE

In this section we use the derivative found in Theorem 3 to characterize distributions that maximize the MMSE. Unlike previous approaches, the approach laid out in this paper is systematic and produces both sufficient and necessary conditions for optimality. Moreover, the approach is fairly general and works for a large class of channels $P_{\mathbf{Y}|\mathbf{X}}$. We begin by introducing necessary mathematical tools.

A. Optimization Theorems

We will need the following optimization theorems.

Theorem 4. (Optimization Theorems [11].)

1) (Extreme Value Theorem.) *For a compact topological space $\mathcal{F}$ and an upper semicontinuous function $f : \mathcal{F} \mapsto \mathbb{R}$*

$$\sup_{F\in\mathcal{F}} f(F) = \max_{F\in\mathcal{F}} f(F). \qquad (8)$$

Moreover, the solution is unique if f is strictly concave.

2) (Necessary Condition for Optimality.) *Let $\mathcal{F}$ be a convex topological space and let $f : \mathcal{F} \mapsto \mathbb{R}$ have a Gâteaux derivative $\Delta_Q f(F)$ as defined in* (4)*. Suppose that $F^\star \in \mathcal{F}$ is a maximizer of f, then*

$$\Delta_Q f(F^\star) \le 0, \forall Q \in \mathcal{F}. \qquad (9)$$

3) (Necessary and Sufficient Condition for Optimality.) *The condition in* (9) *is sufficient if in addition f is concave.*

4) (KKT Conditions.) *Let $\mathcal{F}$ be a convex topological space, and let $f : \mathcal{F} \mapsto \mathbb{R}$ be a concave function on $\mathcal{F}$ and $g : \mathcal{F} \mapsto \mathbb{R}$ a convex function on $\mathcal{F}$. Assume there exists a point $F \in \mathcal{F}$ such that $g(F) < 0$. Let*

$$\mu = \sup_{F\in\mathcal{F} \text{ and } g(F)\le 0} f(F). \qquad (10)$$

Then, there is a constant $\lambda \ge 0$ such that

$$\mu = \sup_{F\in\mathcal{F}} (f(F) - \lambda g(F)). \qquad (11)$$

Furthermore, if the supremum in (10) *is achieved by F_0, it is achieved by F_0 in* (11) *and $\lambda g(F_0) = 0$.*

B. Analytic Functions and the Size of the Uniqueness Set

Part of our analysis will require identifying the sizes of sets on which two analytic functions can agree without being identical everywhere (i.e., uniqueness sets) for which the following theorem will be used.

Theorem 5. (Identity Theorems [12].) *Let $\mathcal{X} \subset \mathbb{R}^n$ and let $f, g : \mathcal{X} \to \mathbb{R}$ be two real-analytic functions on $\mathcal{X}$ that agree on some set $\mathcal{E} \subset \mathcal{X}$. Then, f and g agree on $\mathcal{X}$ if one of the following conditions is satisfied:*

1) $\mathcal{E}$ is an open set;

2) $\mathcal{E}$ is a set of positive Lebesgue measure; or

3) $n = 1$ and $\mathcal{E}$ has a limit point in $\mathcal{X}$.

C. Bounded Input: General Case

In this section we seek to find

$$\sup_{P_{\mathbf{X}}\in\mathcal{F}_\infty(\mathsf{S})} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}).$$

Theorem 6. *For any $P_{\mathbf{Y}|\mathbf{X}}$ and any compact $\mathsf{S} \subset \mathbb{R}^n$*

$$\sup_{P_{\mathbf{X}}\in\mathcal{F}_\infty(\mathsf{S})} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) = \max_{P_{\mathbf{X}}\in\mathcal{F}_\infty(\mathsf{S})} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}). \qquad (12a)$$

Moreover, $P_{\mathbf{X}}^\star$ is an optimal input distribution in (12a) *if and only if for all $Q_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S})$,*

$$\mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) \le \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}). \qquad (12b)$$

Proof. The proof of (12a) follows from the fact that $P_{\mathbf{X}} \mapsto \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}})$ is an upper semicontinuous function, as shown in Theorem 2, using that $\mathcal{F}_\infty(\mathsf{S})$ is a sequentially compact set, as shown in [9, Lemma 1], and applying property 1) from Theorem 4. Finally, the statement in (12b) follows from property 2) and property 3) in Theorem 4, and the derivative expression for the MMSE in Theorem 3. □

In this work we seek to make statements about the size of the support of an optimal input distribution. Therefore, it is convenient to re-write the condition in (12b) in an equivalent form as conditions that involve statements about the support of an optimal input distribution.

Proposition 1. *$P_{\mathbf{X}}^\star$ is an optimal input distribution in* (12a) *if and only if the following two conditions hold:*

1) for all $\mathbf{x} \in \mathsf{S}$

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2|\mathbf{X} = \mathbf{x}\right] \le \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}); \textit{and} \qquad (13a)$$

2) for all $\mathbf{x} \in \mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathsf{S}$

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2|\mathbf{X} = \mathbf{x}\right] = \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}). \qquad (13b)$$

Definition 3. (Dense and Nowhere Dense Sets.)

- *A set $\mathcal{A} \subset \mathcal{X}$ is said to be* dense *in $\mathcal{X}$ if every element $\mathbf{x} \in \mathcal{X}$ either belongs to $\mathcal{A}$ or is a limit point of $\mathcal{A}$.*
- *A set $\mathcal{A} \subset \mathcal{X}$ is said to be* nowhere dense *if, for every nonempty open set $\mathcal{U} \subset \mathcal{X}$, the intersection $\mathcal{U} \cap \mathcal{A}$ is not dense in $\mathcal{U}$.*

Proposition 2. *Suppose that the function*

$$g(\mathbf{x}) \triangleq \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2|\mathbf{X} = \mathbf{x}\right], \qquad (14)$$

is non-constant and real-analytic on S*. Then, an optimal input distribution in* (12a) $P^\star_{\mathbf{X}}$*, satisfies the following:*

- *for* $\mathsf{S} \subset \mathbb{R}^n$ *where* $n \geq 1$*,* $\mathcal{E}(P^\star_{\mathbf{X}})$ *is a nowhere dense set of Lebesgue measure zero; and*
- *for* $\mathsf{S} \subset \mathbb{R}$*,* $\mathcal{E}(P^\star_{\mathbf{X}})$ *has finite cardinality (i.e., an optimal input distribution is discrete with finitely many points).*

Proof. If $P^\star_{\mathbf{X}}$ is a maximizer in (12a), then by (13b)

$$g(\mathbf{x}) = \mathrm{mmse}(P^\star_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}), \forall \mathbf{x} \in \mathcal{E}(P^\star_{\mathbf{X}}). \tag{15}$$

In other words, $g(\mathbf{x})$ is constant on $\mathcal{E}(P^\star_{\mathbf{X}})$.

We first focus on the general case of $n \geq 1$. Towards a contradiction suppose that $\mathcal{E}(P^\star_{\mathbf{X}}) \subseteq \mathsf{S}$ is *not* a nowhere dense set of S. Then there exists some open set $\mathcal{O}$ such that $\mathcal{O} \cap \mathcal{E}(P^\star_{\mathbf{X}})$ is dense in $\mathcal{O}$. Moreover, by (15) $g(\mathbf{x})$ is a constant on $\mathcal{O} \cap \mathcal{E}(P^\star_{\mathbf{X}})$. Since, $g(\mathbf{x})$ is continuous and $\mathcal{O} \cap \mathcal{E}(P^\star_{\mathbf{X}})$ is dense in $\mathcal{O}$ we have that $g(\mathbf{x})$ is constant on $\mathcal{O}$ by the definition of continuity. Finally, since $\mathcal{O}$ is an open set of S by property 1 of Theorem 5 we have that $g(\mathbf{x})$ is constant on all of S. However, this contradicts our assumption that $g(\mathbf{x})$ is non-constant on S and, therefore, $\mathcal{E}(P^\star_{\mathbf{X}})$ is a nowhere dense set.

The conclusion that $\mathcal{E}(P^\star_{\mathbf{X}})$ has Lebesgue measure zero follows by assuming, towards a contradiction, that $\mathcal{E}(P^\star_{\mathbf{X}})$ is a set of positive Lebesque measure. By (15) $g(\mathbf{x})$ is constant on $\mathcal{E}(P^\star_{\mathbf{X}}) \subset \mathsf{S}$ and using Theorem 5 we conclude that $g(\mathbf{x})$ must be constant on S.

The proof in the case of $n = 1$ is relegated to [9]. □

The result of Proposition 2 for $n > 1$ show that the support of an optimal input distribution is small in two ways. First, the support is small in measure theoretic terms and has zero Lebesgue measure. Second, the support is small topologically and is a nowhere dense which loosely speaking implies that the elements of the support are not tightly clustered. An interesting question, which we will address shortly, is whether the size of the support is also small when measured in terms of cardinality. For example, for $n = 1$ we already know that this is the case and the support has finite cardinality. It turns out that in general, for $n > 1$, the support of an optimal distribution might not be of finite or even countably infinite cardinality.

Next, we show that the conditions on $g(\mathbf{x})$ in Proposition 2 are not very restrictive and work for a variety of settings (e.g., Gaussian noise).

Lemma 1. *Let* $P_{\mathbf{Y}|\mathbf{X}}$ *be such that* $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$ *and where* $\mathbf{X}$ *and* $\mathbf{Z}$ *are independent, and suppose that the pdf of* $\mathbf{Z} \sim f_{\mathbf{Z}}(\mathbf{z})$ *is a complex-analytic function on an open subset of* $\mathbb{C}^n$ *that contains* $\mathbb{R}^n$*. Moreover, assume that* $f_{\mathbf{Z}}(\mathbf{z}) > 0$ *for all* $\mathbf{z} \in \mathbb{R}$*. Then,* $g(\mathbf{x})$ *defined in* (14) *is a real analytic function on* $\mathbb{R}^n$*.*

D. Bounded Input: Gaussian Noise Case

This section looks at the case when $P_{\mathbf{Y}|\mathbf{X}}$ is Gaussian.

Proposition 3. (Univariate Gaussian.) *Let* $P_{Y|X}(y|x) = \mathcal{N}(x, 1)$*; then for the optimization problem*

$$\max_{P_X \in \mathcal{F}_\infty([-A,A])} \mathrm{mmse}(P_X, P_{Y|X}), \tag{16}$$

we have the following:

- *an optimal input distribution in* (16) *is discrete with finitely many points. Moreover, the optimizing input distribution is unique and symmetric;*
- $\{\pm A\} \subseteq \mathcal{E}(P^\star_{\mathbf{X}})$ *for every* $A \geq 0$*; and*
- $\{\pm A\} = \mathcal{E}(P^\star_{\mathbf{X}})$ *if and only if* $A \leq \bar{A}_B \approx 1.05647$*.*

For the multivariate case we have the following generalization of Proposition 3.

Proposition 4. (Multivariate Gaussian.) *Let* $P_{\mathbf{Y}|\mathbf{X}} = \mathcal{N}(\mathbf{x}, \mathbf{I})$ *and let* $\mathcal{C}(r) \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = r\}$*. Then, for the optimization problem*

$$\max_{P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathcal{B}_{\mathbf{0}}(R))} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}), \tag{17}$$

we have the following:

- *the optimal input distribution* $P^\star_{\mathbf{X}}$ *is unique and spherically symmetric. Moreover,* $\mathcal{E}(P^\star_{\mathbf{X}}) = \bigcup_{i=1}^N \mathcal{C}(r_i)$*, where* $N < \infty$ *for some* $\{r_i\}_1^N$*;*
- $\mathcal{C}(R) \subseteq \mathcal{E}(P^\star_{\mathbf{X}})$ *for every* $R \geq 0$*; and*
- $\mathcal{C}(R) = \mathcal{E}(P^\star_{\mathbf{X}})$ *if and only if* $R \leq \bar{R} = \Theta(\sqrt{n})$*.*

Note that the result of Proposition 4 shows that an optimal input distribution can be supported on the set $\mathcal{C}(R)$ which is a nowhere dense set of Lebesgue measure zero. However, note that the set $\mathcal{C}(R)$ does have an uncountably infinite cardinality. Hence, for $n > 1$ the conclusion in Proposition 2 is not superfluous and in general cannot be strengthened, and discrete inputs are in general not optimal for $n > 1$. However, do note that the number of possible spheres that make up $\mathcal{E}(P^\star_{\mathbf{X}})$ is finite. In other words, the magnitude $\|\mathbf{X}\|$ is a discrete random variable with finitely many points.

In Proposition 4, the constant that determines $\bar{R}$ can be difficult to evaluate, but it can be shown that it is sufficient to take $\bar{R} \leq \sqrt{n}$. Proposition 4 can be used to find the capacity of a MIMO channel given an amplitude constraint.

Proposition 5. (Amplitude Constrained MIMO.) *For*

$$\max_{\mathbf{X}: \mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}), \text{ where } \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{18}$$

the optimal input distribution is uniform on the set $\mathcal{C}(R) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = R\}$ *(i.e., boundary of the ball) if* $R \leq \sqrt{n}$*.*

Note that Proposition 5 establishes capacity in the small amplitude regime (i.e., $R \leq \sqrt{n}$) in the massive MIMO case (i.e., the number of antennas approaches infinity) [13]. A similar argument can also be applied to the MIMO wiretap channel.

E. Bounded Input: Poisson Noise Case

The Poisson random transformation is governed by the following conditional distribution:

$$p_{Y|X}(y|x) = \frac{1}{y!} x^y \mathrm{e}^{-x}, \, x \geq 0, \, y = 0, 1, ..., \tag{19}$$

where we use the convention that $0^0 = 1$. It is well known that the conditional expectation is given by

$$\mathbb{E}[X|Y = y] = \frac{(y+1) p_Y(y+1; P_X)}{p_Y(y; P_X)}, \, y = 0, 1, ..., \tag{20}$$

where $p_Y(y; P_X)$ is the marginal probability mass function (pmf) of Y induced by input distribution P_X.

The following theorem characterizes the structure of a least favorable prior for the Poisson case.

Proposition 6. (Poisson Noise Case.) *Let $P_{Y|X}$ be as in* (19)*. Then, for the optimization problem*

$$\max_{P_X \in \mathcal{F}_\infty([0,A])} \operatorname{mmse}(P_X, P_{Y|X}), \tag{21}$$

we have the following:

- *an optimal input distribution in* (21) *is discrete with finitely many points; and*
- *$\mathcal{E}(P_X^\star) = \{0, A\}$ if and only if $A \le \bar{A} \approx 0.9129$ where $\bar{A}$ is the solution of the equation $2\mathrm{e}^{\frac{x}{2}}(x-1) + x\mathrm{e}^x - 2 = 0$ for $x > 0$. Moreover, the optimal probability assignment is given by $P_X^\star[X=0] = \frac{1}{1+\mathrm{e}^{\frac{A}{2}}}$, and the MMSE is given by*
$$\operatorname{mmse}(P_X^\star, P_{Y|X}) = A^2 \left(P_X^\star[X=0]\right)^2.$$

F. Generalized Input Moment Constraints

In this section we seek to find

$$\sup_{P_{\mathbf{X}} \in \mathcal{F}(f;\alpha)} \operatorname{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \tag{22a}$$

$$\text{where } \mathcal{F}(f;\alpha) = \{P_{\mathbf{X}} : \mathbb{E}_{P_{\mathbf{X}}}[f(\mathbf{X})] \le \alpha\}. \tag{22b}$$

for some given $f : \mathbb{R}^n \to \mathbb{R}$ independent of $P_{\mathbf{X}}$. Observe that the set $\mathcal{F}(f;\alpha)$ is convex. In addition, we assume that $f(\mathbf{X})$ is a non-negative monotonically increasing function of $\|\mathbf{X}\|$ which by the Markov inequality and Prokhorov theorem implies that $\mathcal{F}(f;\alpha)$ is a sequentially compact set. An example of an $f(\cdot)$ that satisfies such a condition is $f(\mathbf{X}) = \|\mathbf{X}\|^r$ for any $r > 0$.

Theorem 7. *Suppose the MMSE in the optimization problem in* (22) *is an upper semicontinuous function. Then, the supremum in* (22a) *is attainable by some input distribution $P_{\mathbf{X}}^\star$. Moreover, $P_{\mathbf{X}}^\star$ is optimal if and only if the following two conditions hold:*

1) for all $\mathbf{x} \in \mathbb{R}^n$

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] - \lambda\left(f(\mathbf{x}) - \alpha\right) \le \operatorname{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}); \textit{and}$$

2) for all $\mathbf{x} \in \mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathbb{R}^n$

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] - \lambda\left(f(\mathbf{x}) - \alpha\right) = \operatorname{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}).$$

Next, we look at the special case of multivariate Gaussian.

Proposition 7. *Let $P_{\mathbf{Y}|\mathbf{X}} = \mathcal{N}(\mathbf{x}, \mathbf{I})$. Then for the optimization problem in* (22) *we have the following:*

- *the optimal input distribution is unique and symmetric.*
- *if $f(\mathbf{x}) = \omega\left(\|\mathbf{x}\|^2\right)$, then the support of the optimal input distribution is bounded (i.e., $\mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathcal{B}_{\mathbf{0}}(R)$ for some $R > 0$);*
- *if $f(\mathbf{x}) = \|\mathbf{x}\|^2$, then the optimal input distribution is given by $\mathbf{X} \sim \mathcal{N}(0, \alpha I)$; and*
- *if $f(\mathbf{x}) = o(\|\mathbf{x}\|^2)$, then the support of the optimal input distribution is unbounded (i.e., there is no $R \ge 0$ such that $\mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathcal{B}_{\mathbf{0}}(R)$).*

It is important to point out that the proof of the case $f(\mathbf{x}) = o(\|\mathbf{x}\|^2)$ in Proposition 7 does not require the assumption that $P_{\mathbf{Y}|\mathbf{X}}$ is Gaussian, and holds under the general assumptions of Theorem 7.

Observe that according to Proposition 7, in the case of $f(\mathbf{x}) = \omega\left(\|\mathbf{x}\|^2\right)$, the optimal distribution has a bounded support and, therefore, from Proposition 2 we have that the support is a nowhere dense set of Lebesgue measure zero.

V. Conclusion

In this work we have examined the structure of the support of least favorable prior distributions. We have shown that, under some mild conditions, the support of a least favorable distribution must be a nowhere dense set of Lebesgue measure zero. Our results also produce necessary and sufficient conditions for optimality and, in most cases, can be easily evaluated as has been demonstrated by the Gaussian and the Poisson examples. An interesting future direction is to consider the problem where for $\lambda \ge 0$ we seek to maximize

$$\max_{P_{\mathbf{X}}} \left(\operatorname{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) - \lambda \operatorname{mmse}(P_{\mathbf{X}}, Q_{\mathbf{Y}|\mathbf{X}})\right).$$

For example, taking $P_{\mathbf{Y}|\mathbf{X}} = \mathcal{N}(\mathbf{H}\mathbf{x}, \mathbf{I})$ and $Q_{\mathbf{Y}|\mathbf{X}} = \mathcal{N}(\mathbf{H}_0\mathbf{x}, \mathbf{I})$ might potentially generalize the *single crossing point property*, shown in [14] and discussed in great detail in [15] and [16].

References

[1] J. C. Berry, "Minimax estimation of a bounded normal mean vector," *Journal of Multivariate Analysis*, vol. 35, no. 1, pp. 130–139, 1990.

[2] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[3] M. Raginsky, "On the information capacity of Gaussian channels under small peak power constraints," in *46th Annual Proc. Allerton Conf. Commun., Control and Comp.*, Monticello, IL, USA, 2008, pp. 286–293.

[4] J. G. Smith, "The information capacity of amplitude-and variance-constrained scalar Gaussian channels," *Information and Control*, vol. 18, no. 3, pp. 203–219, 1971.

[5] J. Fahs and I. Abou-Faycal, "On properties of the support of capacity-achieving distributions for additive noise channel models with input cost constraints," *IEEE Trans. Inf. Theory*, 2017, to appear.

[6] M. Ghosh, "Uniform approximation of minimax point estimates," *The Annals of Mathematical Statistics*, pp. 1031–1047, 1964.

[7] G. Casella and W. E. Strawderman, "Estimating a bounded normal mean," *The Annals of Statistics*, pp. 870–878, 1981.

[8] E. Marchand and W. E. Strawderman, "Estimation in restricted parameter spaces: A review," *Lecture Notes-Monograph Series*, pp. 21–44, 2004.

[9] A. Dytso, R. Bustin, H. V. Poor, and S. Shamai (Shitz), "On the structure of the least favorable prior distributions," 2018. [Online]. Available: http://www.princeton.edu/%7Eadytso/papers/LFDforMMSE.pdf

[10] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1289–1301, 2012.

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

[12] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions.* Springer Science & Business Media, 2002.

[13] A. Dytso, M. Goldenbaum, H. V. Poor, and S. Shamai (Shitz), "A generalized Ozarow-Wyner capacity bound with applications," in *Proc. IEEE Int. Symp. Inf. Theory*, Monticello, IL, USA, 2017, pp. 1058–1062.

[14] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, 2011.

[15] R. Bustin, M. Payaró, D. P. Palomar, and S. Shamai, "On MMSE crossing properties and implications in parallel vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 818–844, 2013.

[16] A. Dytso, R. Bustin, H. V. Poor, and S. Shamai (Shitz), "A view of information-estimation relations in Gaussian networks," *Entropy*, vol. 19, no. 8, 2017.