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 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 some mild regularity conditions, it is shown that the support of the lease favorable distributions must necessarily be very small and is contained in a nowhere dense set of Lebesgue measure zero. The results of this paper produces both sufficient and necessary conditions for optimality, do not rely on Gaussian statistics assumption, and are not sensitive to the dimensionality of random vectors. The results are evaluate for the univariate and multivariate random Gaussian cases, and the Poisson case. Finally, as one of the applications, we show how our result can be used to characterize capacity of Gaussian MIMO channels with an amplitude constraint.

I. Introduction

The *minimum mean square error* (MMSE) of 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 types of 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 distribution 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 we take in this work is based on the theory of convex optimization and allows us to produce systematic and very 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], while not explicitly stated, has shown that, with the support constraint, the least favorable priors distribution 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 sufficient condition for the optimality of a three point prior. In contrast, the methodology used in this paper produces both sufficient and necessary condition 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

*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 National Science Foundation under Grants CCF-1420575 and CNS-1456793. The work of S. Shamai and R. Bustin was supported by the Unions Horizon 2020 Research and Innovation Programme Grant 694630. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies.

mean the interested reader is referred to [8] and reference therein.

B. Outline and Paper Contributions

Our contributions are as follows:

1) In Section II we review important properties of the MMSE needed in our analysis.
2) 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 at the distribution $P_{\mathbf{X}}$ in the direction of the distribution $Q_{\mathbf{X}}$.
3) In Section IV we apply theory of convex optimization to maximize the MMSE and present:
   - In Section IV-A, Theorem 4 presents required theorems from convex optimization;
   - In Section IV-B, Theorem 5 presents required theorems of analytic functions;
   - In Section IV-C looks at the case of the compact support constraint. In Theorem 6 it is shown that a least favorable input distribution exists for an arbitrary $P_{\mathbf{Y}|\mathbf{X}}$ and derived necessary and sufficient conditions for the optimality. Moreover, Proposition 1, under some mild conditions, characterizes the structure of the support of least favorable prior distribution and shows that the support must be a *nowhere dense set of Lebesgue measure zero*;
   - In Section IV-D, 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 capacity of MIMO channels. Surprisingly, Proposition 5 also characterize capacity of the MIMO amplitude channel in a regime where the number of antennas goes to infinity;
   - In Section IV-E, Proposition 6 looks at the Poisson noise case; and
   - Section IV-F, looks at least favorable priors under the generalized moment constraints.
4) Section VI 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;
- We denote an n-dimensional ball of radius R centered at $\mathbf{0}$ as $\mathcal{B}_{\mathbf{0}}(R) \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq R\}$;
- 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 possible 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 overview some important properties of the MMSE.

A. Fundamental Theorems of the MMSE Estimation

Theorem 1. (Fundamental Theorems of the MMSE Estimation.)

1) (Orthogonality Principle.) *For any $f : \mathbb{R}^n \to \mathbb{R}^k$ such that $f \in L^2$*

$$\mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^T f(\mathbf{Y})] = 0. \tag{2a}$$

2) (Pythagorean Theorem.) *For any $f : \mathbb{R}^n \to \mathbb{R}^k$ such that $f \in L^2$*

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right] &= \mathbb{E}\left[\|\mathbf{X} - f(\mathbf{Y})\|^2\right] \\ &\quad - \mathbb{E}\left[\|f(\mathbf{Y}) - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right].\end{aligned} \tag{2b}$$

3) (Conditional Expectation is the Optimal Estimator.)

$$\begin{aligned}&\operatorname{mmse}(\mathbf{X}|\mathbf{Y}) \\ &= \inf_{f(\cdot): f \text{ is measurable and } f \in L^2} \mathbb{E}\left[\|\mathbf{X} - f(\mathbf{Y})\|^2\right] \\ &= \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right].\end{aligned} \tag{2c}$$

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

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

Continuity properties of the MMSE would 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-semicontinuous) at a point $x_0 \in \mathcal{F}$ if*

$$\limsup_{x \to x_0} f(x) \leq f(x_0) \quad \left(\textit{resp. } \liminf_{x \to x_0} f(x) \geq f(x_0)\right). \tag{4}$$

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

We summaries operator properties of the MMSE in the next theorem.

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*

$$\mathcal{M}(\mathsf{S}) = \{P_{\mathbf{XY}} : \forall P_{\mathbf{Y}|\mathbf{X}} \text{ and } P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S}) \text{ where } \mathsf{S} \text{ is bounded}\}. \tag{5}$$

- *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 theGâteaux derivative of a function* $g : \mathcal{F} \to \mathbb{R}$ *at* P *in the direction of* Q *as*

$$\Delta_Q g(P) \triangleq \lim_{\lambda \to 0} \frac{g\left((1-\lambda)P + \lambda Q\right) - g\left(P\right)}{\lambda}. \tag{6}$$

The Gâteaux derivative is simply a 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* $P_{\mathbf{Y}|\mathbf{X}}$ *the Gâteaux derivative of the MMSE is given by*

$$\begin{aligned}&\Delta_{Q_{\mathbf{X}}} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\ &= \mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) - \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}),\end{aligned} \tag{7a}$$

where

$$\mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \triangleq \mathbb{E}_{Q_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right]. \tag{7b}$$

Proof: Let $P_\lambda = (1-\lambda)P_{\mathbf{X}} + \lambda Q_{\mathbf{X}}$. From the definition of Gâteaux derivative in (6) we have to look at

$$\begin{aligned}&\mathrm{mmse}(P_\lambda, P_{\mathbf{Y}|\mathbf{X}}) - \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\ &= \mathbb{E}_{P_\lambda}\left[\|\mathbf{X} - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right] - \mathbb{E}_{P_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\ &\overset{a)}{=} \mathbb{E}_{P_\lambda}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] - \mathbb{E}_{P_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\ &\quad - \mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\ &\overset{b)}{=} (1-\lambda)\mathbb{E}_{P_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\ &\quad + \lambda\mathbb{E}_{Q_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] - \mathbb{E}_{P_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\ &\quad - \mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\ &= \lambda\mathbb{E}_{Q_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] - \lambda\mathbb{E}_{P_{\mathbf{X}}}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2\right] \\ &\quad - \mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right],\end{aligned} \tag{8}$$

where the steps follow from: a) Pythagorean identity in (2b); and b) using the property that expected value is a linear operator on a set of distributions. Next by dividing (8) by λ and taking $\lambda \to 0$ we have that

$$\begin{aligned}&\Delta_{Q_{\mathbf{X}}} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\ &= \mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) - \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \\ &\quad - \lim_{\lambda \to 0} \frac{\mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right]}{\lambda},\end{aligned} \tag{9}$$

Next, we show that the third term in (9) is zero or that

$$\lim_{\lambda \to 0} \frac{\mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right]}{\lambda} = 0. \tag{10}$$

In order to simplify the notation, we assume that $P_{\mathbf{Y}|\mathbf{X}}$ is an absolutely continuous distribution $P_{\mathbf{Y}|\mathbf{X}}$ with a density $f_{\mathbf{Y}|\mathbf{X}}$ and the conditional expectation can be written as

$$\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] = \frac{\mathbb{E}_{P_{\mathbf{X}}}\left[\mathbf{X} f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{X})\right]}{f_{\mathbf{Y}}(\mathbf{y}; P_{\mathbf{X}})} \triangleq \frac{q(\mathbf{y}; P_{\mathbf{X}})}{f_{\mathbf{Y}}(\mathbf{y}; P_{\mathbf{X}})},$$

where $f_{\mathbf{Y}}(\mathbf{y}; P_{\mathbf{X}})$ is an output distribution induced by the input $\mathbf{X} \sim P_{\mathbf{X}}$ that is

$$f_{\mathbf{Y}}(\mathbf{y}; P_{\mathbf{X}}) = \int f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \mathrm{d}P_{\mathbf{X}}(\mathbf{x}).$$

Next, we re-write (9) as follows:

$$\begin{aligned}&\frac{\mathbb{E}_{P_\lambda}\left[\|\mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}]\|^2\right]}{\lambda} \\ &= \frac{1}{\lambda}\mathbb{E}_{P_\lambda}\left[\left\|\frac{q(\mathbf{Y}; P_{\mathbf{X}})}{f_{\mathbf{Y}}(\mathbf{Y}; P_{\mathbf{X}})} - \frac{q(\mathbf{Y}; P_\lambda)}{f_{\mathbf{Y}}(\mathbf{Y}; P_\lambda)}\right\|^2\right] \\ &= \frac{\mathbb{E}_{P_\lambda}\left[\left\|\frac{q(\mathbf{Y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y};P_\lambda) - q(\mathbf{Y};P_\lambda) f_{\mathbf{Y}}(\mathbf{Y};P_{\mathbf{X}})}{f_{\mathbf{Y}}(\mathbf{Y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y};P_\lambda)}\right\|^2\right]}{\lambda} \\ &\overset{a)}{=} \frac{\mathbb{E}_{P_\lambda}\left[\left\|\frac{\lambda(q(\mathbf{Y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y};Q_{\mathbf{X}}) - q(\mathbf{Y};Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y};P_{\mathbf{X}}))}{f_{\mathbf{Y}}(\mathbf{Y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y};P_\lambda)}\right\|^2\right]}{\lambda} \\ &\overset{b)}{=} \frac{\lambda^2\mathbb{E}_{P_\lambda}\left[\frac{\|q(\mathbf{Y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y};Q_{\mathbf{X}}) - q(\mathbf{Y};Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y};P_{\mathbf{X}})\|^2}{f_{\mathbf{Y}}^2(\mathbf{Y};P_{\mathbf{X}}) f_{\mathbf{Y}}^2(\mathbf{Y};P_\lambda)}\right]}{\lambda} \\ &= \lambda\int \frac{\|q(\mathbf{y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};Q_{\mathbf{X}}) - q(\mathbf{y};Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}})\|^2}{f_{\mathbf{Y}}^2(\mathbf{y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};P_\lambda)} \mathrm{d}\mathbf{y},\end{aligned} \tag{11}$$

where the steps follow from: a) using that $f_{\mathbf{Y}}(\mathbf{Y}; P_\lambda) = (1-\lambda) f_{\mathbf{Y}}(\mathbf{Y}; P_{\mathbf{X}}) + \lambda f_{\mathbf{Y}}(\mathbf{Y}; Q_{\mathbf{X}})$ and $q(\mathbf{Y}; P_\lambda) = (1-\lambda) q(\mathbf{Y}; P_{\mathbf{X}}) + \lambda q(\mathbf{Y}; Q_{\mathbf{X}})$ which leads to

$$\begin{aligned}&q(\mathbf{Y}; P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y}; P_\lambda) - q(\mathbf{Y}; P_\lambda) f_{\mathbf{Y}}(\mathbf{Y}; P_{\mathbf{X}}) \\ &= (1-\lambda) q(\mathbf{Y}; P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y}; P_{\mathbf{X}}) + \lambda q(\mathbf{Y}; P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y}; Q_{\mathbf{X}}) \\ &\quad - (1-\lambda) q(\mathbf{Y}; P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y}; P_{\mathbf{X}}) - \lambda q(\mathbf{Y}; Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y}; P_{\mathbf{X}}) \\ &= \lambda(q(\mathbf{Y}; P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y}; Q_{\mathbf{X}}) - q(\mathbf{Y}; Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{Y}; P_{\mathbf{X}})); \text{ and}\end{aligned}$$

b) using the absolute scalability of the Euclidian norm. Next observe that

$$g_\lambda(\mathbf{y}) \triangleq \frac{\lambda \left\| q(\mathbf{y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};Q_{\mathbf{X}}) - q(\mathbf{y};Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}}) \right\|^2}{f^2_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};P_\lambda)}$$
$$= \frac{\left\| q(\mathbf{y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};Q_{\mathbf{X}}) - q(\mathbf{y};Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}}) \right\|^2}{f^2_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}})} \cdot \frac{\lambda}{(1-\lambda) f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}}) + \lambda f_{\mathbf{Y}}(\mathbf{y};Q_{\mathbf{X}})},$$

is a monotonically increasing function of $\lambda \in (0,1)$ for every $\mathbf{y}$ since

$$\frac{d}{d\lambda} \frac{\lambda}{(1-\lambda) f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}}) + \lambda f_{\mathbf{Y}}(\mathbf{y};Q_{\mathbf{X}})} = \frac{f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}})}{\left(\lambda f_{\mathbf{Y}}(\mathbf{y};Q_{\mathbf{X}}) + (1-\lambda) f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}})\right)^2} \geq 0.$$

Therefore, by the monotone convergence theorem we can exchange the limit and the expectation

$$\lim_{\lambda \to 0} \frac{\mathbb{E}_{P_\lambda}\left[\left\| \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}] - \mathbb{E}_{P_\lambda}[\mathbf{X}|\mathbf{Y}] \right\|^2 \right]}{\lambda}$$
$$= \int \lim_{\lambda \to 0} \frac{\lambda \left\| q(\mathbf{y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};Q_{\mathbf{X}}) - q(\mathbf{y};Q_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}}) \right\|^2}{f^2_{\mathbf{Y}}(\mathbf{y};P_{\mathbf{X}}) f_{\mathbf{Y}}(\mathbf{y};P_\lambda)} \mathrm{d}\mathbf{y}$$
$$= 0. \tag{12}$$

This concludes the proof for the case when $P_{\mathbf{Y}|\mathbf{X}}$ is an absolutely continuous distribution. Evidently, the case when when $P_{\mathbf{Y}|\mathbf{X}}$ does not have a pdf or pmf is handle in above by replacing $f_{\mathbf{Y}|\mathbf{X}}$ with $\mathrm{d}P_{\mathbf{Y}|\mathbf{X}}$. This concludes the proof. ∎

It is interesting to note that Theorem 3 holds with no assumption on the distribution $P_{\mathbf{X}}, Q_{\mathbf{X}}$ and $P_{\mathbf{Y}|\mathbf{X}}$. This is because the interchange of limit and expectation in (12) was done by using the monotone convergence theorem instead of more ubiquitous dominate convergence theorem.

IV. Optimization of the MMSE

In this section we apply the directional derivative found in Theorem 3 to characterize distributions that maximize the MMSE. Unlike the 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 theorem in our analysis.

Theorem 4. (Optimization Theorems.)

1) (Extreme Value Theorem [11, Section 2.13].) *For any compact topological space $\mathcal{F}$ and any upper semicontinuous function $f : \mathcal{F} \mapsto \mathbb{R}$*

$$\sup_{F \in \mathcal{F}} f(F) = \max_{F \in \mathcal{F}} f(F). \tag{13}$$

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

2) (Necessary Condition for Optimality [11, Section 7.4].) *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* (6)*. Suppose that $F^\star \in \mathcal{F}$ is a maximizer of f, then*

$$\Delta_Q f(F^\star) \leq 0, \, \forall Q \in \mathcal{F}. \tag{14}$$

3) (Necessary and Sufficient Condition for Optimality). *The condition in* (14) *is also sufficient if in addition the function f is concave on $\mathcal{F}$.*

4) (KKT Conditions [11, Section 8.3].) *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*

$$\mu = \sup_{F \in \mathcal{F} \text{ and } g(F) \leq 0} f(F). \tag{15}$$

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

$$\mu = \sup_{F \in \mathcal{F}} \left(f(F) - \lambda g(F) \right). \tag{16}$$

Furthermore, if the supremum in (15) *is achieved by F_0, it is achieved by F_0 in* (16) *and*

$$\lambda g(F_0) = 0. \tag{17}$$

B. Analytic Function and the Size of the Uniqueness Set

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

Theorem 5. (Identity Theorem for Real-Analytic Functions [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

$$\sup_{P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S})} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}). \tag{18}$$

In order to apply the optimization theorems, summarized in Theorem 4, we will need the following result about the properties of $\mathcal{F}_\infty(\mathsf{S})$.

Lemma 1. *For a compact $\mathsf{S} \subset \mathbb{R}^n$ the set $\mathcal{F}_\infty(\mathsf{S})$ is sequentially compact.*

Proof: Note that since all probability measure are supported on a compact set, they are automatically uniformly tight. Therefore, by Prohorov's theorem [13] $\mathcal{F}_\infty(\mathsf{S})$ is sequentially compact. ∎

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

$$\sup_{P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S})} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) = \max_{P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S})} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}). \tag{19a}$$

Moreover, let $P_{\mathbf{X}}^\star$ be an optimal input distribution in (19a)*, then a necessary and sufficient condition for the optimality of $P_{\mathbf{X}}^\star$ is given by*

$$\mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) \leq \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}),\ \forall Q_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S}). \tag{19b}$$

Proof: The proof of (19a) follows from using 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, and using that $\mathcal{F}_\infty(\mathsf{S})$ is a sequentiall compact set, as shown in Lemma 1, and applying property 1) from Theorem 4. Finally, the statement in (19b) follows from property 2) and property 3) in Theorem 4, and where the derivative expression of the MMSE in Theorem 3. ∎

In this work we seek to make statements about the size of the support of the optimal input distribution. Therefore, it is convenient to re-write the condition in (19b) in an equivalent term as conditions that involve statements about the support of the optimal input distribution.

Proposition 1. *$P_{\mathbf{X}}^\star$ is an optimal input distribution in* (19a) *if and only if the following two conditions hold:*

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

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \leq \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}); \text{and} \tag{20a}$$

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

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] = \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}). \tag{20b}$$

Proof: To show that (20) implies (19b) simply take the expected value in (20) with respect to $Q_{\mathbf{X}}$. Next, we show that (19b) implies (20).

Assume that $P_{\mathbf{X}}^\star$ is an optimal input distribution. Towards a contradiction we assume that condition (20a) does not hold which implies that there exists some $\mathbf{x}_1 \in \mathsf{S}$ such that

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}_1\right] > \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}). \tag{21}$$

Next let $Q_{\mathbf{X}} = \delta_{\mathbf{x}_1}$ and by using (19b) we have that

$$\begin{aligned}\mathrm{mmse}_{\delta_{\mathbf{x}_1}}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) &= \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}_1\right] \\ &\leq \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}).\end{aligned} \tag{22}$$

Clearly, (22) contradicts (21). The contradiction implies that (20a) is a valid condition.

Now, towards a contradiction suppose that (20b) is not true. By the condition in (20a) (which we just verified) and the assumption that (20b) is not true, there exists a set $\mathsf{S}_1 \subseteq \mathcal{E}(P_{\mathbf{X}}^\star)$ of positive measure (i.e., $P_{\mathbf{X}}^\star(\mathsf{S}_1) = \epsilon > 0$) such that

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] < \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}), \quad \mathbf{x} \in \mathsf{S}_1 \subseteq \mathcal{E}(P_{\mathbf{X}}^\star), \tag{23a}$$

and

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] = \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}), \quad \mathbf{x} \in \mathcal{E}(P_{\mathbf{X}}^\star) \setminus \mathsf{S}_1. \tag{23b}$$

Clearly, we have that

$$\begin{aligned}&\mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) \\ &= \int_{\mathbb{R}} \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \mathrm{d}P_{\mathbf{X}}^\star(\mathbf{x}) \\ &\overset{a)}{=} \int_{\mathcal{E}(P_{\mathbf{X}}^\star)} \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \mathrm{d}P_{\mathbf{X}}^\star(\mathbf{x}) \\ &= \int_{\mathcal{E}(P_{\mathbf{X}}^\star)\setminus\mathsf{S}_1} \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \mathrm{d}P_{\mathbf{X}}^\star(\mathbf{x}) \\ &+ \int_{\mathsf{S}_1} \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \mathrm{d}P_{\mathbf{X}}^\star(\mathbf{x}) \\ &\overset{b)}{=} (1-\epsilon)\mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) \\ &+ \int_{\mathsf{S}_1} \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \mathrm{d}P_{\mathbf{X}}^\star(\mathbf{x}) \\ &\overset{c)}{<} (1-\epsilon)\mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) + \epsilon\,\mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) \\ &= \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}),\end{aligned} \tag{24}$$

where (in)-equalities follow from: a) using the property that $P_{\mathbf{X}}^\star(\mathbb{R}^n) = P_{\mathbf{X}}^\star(\mathcal{E}(P_{\mathbf{X}}^\star)) = 1$; b) using condition (23b); and c) using condition (23a).

Clearly, (24) leads to a contradiction and, therefore, (20b) is a valid condition. This concludes the proof. ∎

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*

$$g(\mathbf{x}) \triangleq \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right], \tag{25}$$

satisfies the following two conditions for all $P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathsf{S})$:

1) $g(\mathbf{x})$ is non-constant on S; and

2) $g(\mathbf{x})$ is a real-analytic function on S.

Then, the optimal input distribution in (19a) *$P_{\mathbf{X}}^\star$ satisfies the following properties:*

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

Proof: If $P_{\mathbf{X}}^\star$ achieves the maximum in (19a), then according to (20b)

$$g(\mathbf{x}) = \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}), \forall \mathbf{x} \in \mathcal{E}(P_{\mathbf{X}}^\star). \tag{26}$$

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

We first focus on the case of $n > 1$. Now towards a contradiction suppose that $\mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathsf{S}$ 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 (26) $g(\mathbf{x})$ 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 on $\mathcal{O}$ 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 a Lebesgue measure zero follows by assuming, towards a contradiction, that $\mathcal{E}(P^\star_{\mathbf{X}})$ is a set of positive Lebesque measure. By (26) $g(\mathbf{x})$ 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.

Next, for the case of $n=1$. Assume that $\mathcal{E}(P^\star_{\mathbf{X}})$ has an infinite cardinality. Then by the Bolzano-Weierstrass theorem there exists a subsequence on $\mathcal{E}(P^\star_{\mathbf{X}})$ that has a limit point in S. Therefore, by property 3) of Theorem 5 the $g(\mathbf{x})$ is a constant function on S. However, this contradicts our assumption that $g(\mathbf{x})$ is non-constant on S. This concludes the proof. ∎

The result of Proposition 2, for $n>1$ show that the support of the optimal input distribution is small in two ways. First, the support is small in terms of measure theory 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 optimal support 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 2. *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 functions 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* (25) *is a real analytic function on $\mathbb{R}^n$.*

D. Bounded Input: Gaussian Noise Case

In this section we look at the case when $P_{\mathbf{Y}|\mathbf{X}}$ is Gaussian. We also need the following results.

Lemma 3. (Slope of the Optimal Estimator.) *Let $|X|\le A$ and $P_{Y|X}=\mathcal{N}(x,1)$. Then,*

$$\frac{d}{dy}\mathbb{E}[X|Y=y]=\mathrm{Var}(X|Y=y),\ \forall y\in\mathbb{R}, \tag{27a}$$

and

$$\max_{X:|X|\le A}\mathrm{Var}(X|Y=y)\le\frac{A^2}{1+A^2},\ \forall y\in\mathbb{R}. \tag{27b}$$

Proof: The identity in (27a) is well know in the literature, for example see [15]. To show (27b) observe that

$$\begin{aligned}&\max_{X:|X|\le A}\mathrm{Var}(X|Y=y)\\&=\max_{X:|X|\le A\text{ and }\mathbb{E}[X^2]\le A^2}\mathrm{Var}(X|Y=y)\\&\le\max_{X:\mathbb{E}[X^2]\le A^2}\mathrm{Var}(X|Y=y).\end{aligned} \tag{28}$$

Next, by noting that the conditional expectation is a minimizer of the conditional variance for every y, that is

$$\begin{aligned}\mathrm{Var}(X|Y=y)&=\mathbb{E}\left[(X-\mathbb{E}[X|Y])^2|Y=y\right]\\&=\inf_{f(\cdot)}\mathbb{E}[(X-f(Y))^2|Y=y]\\&\le\mathbb{E}\left[\left(X-\frac{A^2}{1+A^2}Y\right)^2|Y=y\right],\end{aligned} \tag{29}$$

where the last upper bound follows by choosing possibly sub-optimal function $f(y)=\frac{A^2}{1+A^2}y$. It is not difficult to show that the bound in (29) can only be achieved by $X_G\sim\mathcal{N}(0,A^2)$. This implies that

$$\begin{aligned}\max_{X:\mathbb{E}[X^2]\le A^2}\mathrm{Var}(X|Y=y)&=\mathbb{E}\left[\left(X_G-\frac{A^2Y}{1+A^2}\right)^2|Y=y\right]\\&=\frac{A^2}{1+A^2},\end{aligned} \tag{30}$$

for all $y\in\mathbb{R}$. Combining (28) and (30) concludes the proof of (27b). ∎

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

$$\max_{P_X\in\mathcal{F}_\infty([-A,A])}\mathrm{mmse}(P_X,P_{Y|X}), \tag{31}$$

we have the following:

- *The optimal input distribution in* (31) *is discrete with finitely many points. Moreover, the the optimizing input distribution is unique and symmetric;*
- *The maximizing input distribution always contains mass points at $\{\pm A\}$; and*
- *A random variable $X=\{\pm A\}$ equally likely is optimal if and only if $A\le\bar{A}_B\approx 1.05647$.*

Proof: The fact that an optimal input distribution has finitely many points follows from Proposition 2 and Lemma 2. The uniqueness of the optimal input distribution follows from the fact that for the Gaussian noise case the MMSE is a strictly concave function [10]. The symmetry of the distribution follows from the symmetry of the Gaussian noise. Therefore, in the remaining of this proof we assume that all of the random variables have symmetric distributions.

To show that there must always be mass points at $\pm A$ we first find and bound the derivative of $g(x)$ as follows:

$$\begin{aligned}
&\frac{dg(x)}{dx}\\
&= \frac{d}{dx}\mathbb{E}\left[(X-\mathbb{E}[X|Y])^2|X=x\right]\\
&= \frac{d}{dx}\mathbb{E}\left[(x-\mathbb{E}[X|Y=Z+x])^2\right]\\
&\overset{a)}{=} \mathbb{E}\left[2(x-\mathbb{E}[X|Y=Z+x])(1-\mathrm{Var}(X|Y=Z+x))\right]\\
&= \mathbb{E}\left[2(X-\mathbb{E}[X|Y])(1-\mathrm{Var}(X|Y))|X=x\right]\\
&= \mathbb{E}\left[2(X-\mathbb{E}[X|Y])|X=x\right]\\
&- \mathbb{E}\left[2(X-\mathbb{E}[X|Y])\mathrm{Var}(X|Y)|X=x\right]\\
&\overset{b)}{\geq} 2\mathbb{E}\left[(X-\mathbb{E}[X|Y])|X=x\right]\\
&- 2\mathbb{E}\left[|X-\mathbb{E}[X|Y]|\,|X=x\right]\cdot\sup_{y\in\mathbb{R}}\mathrm{Var}(X|Y=y)\\
&\overset{c)}{\geq} 2\mathbb{E}\left[(X-\mathbb{E}[X|Y])|X=x\right]\\
&- 2\mathbb{E}\left[|X-\mathbb{E}[X|Y]|\,|X=x\right]\cdot\frac{A^2}{1+A^2},
\end{aligned}\tag{32}$$

where (in)-equalities follow from: a) using the identity $\frac{d}{dy}\mathbb{E}[X|Y=y]=\mathrm{Var}(X|Y=y)$ in (27a); b) using modulus inequality and bounding $\mathrm{Var}(X|Y)$; and c) using the bound in (30) that $\sup_{y\in\mathbb{R}}\mathrm{Var}(X|Y=y)\leq\frac{A^2}{1+A^2}$ for all $|X|\leq A$.

Next, we show that the derivative of $g(x)$ around $x=A$ is strictly positive. We have that

$$\begin{aligned}
\frac{dg(x)}{dx}\Big|_{x=A} &\overset{a)}{\geq} 2\mathbb{E}\left[(X-\mathbb{E}[X|Y])|X=A\right]\\
&\quad- 2\mathbb{E}\left[|X-\mathbb{E}[X|Y]|\,|X=A\right]\cdot\frac{A^2}{1+A^2},\\
&\overset{b}{=} \frac{2}{1+A^2}\mathbb{E}\left[(A-\mathbb{E}[X|Y])|X=A\right]\\
&\overset{c)}{\geq} \frac{2}{1+A^2}\mathbb{E}\left[(A-A\cdot 1_{\{Y\geq 0\}}(Y)|X=A\right]\\
&= \frac{2A}{1+A^2}\left(1-\mathbb{P}[Y\geq 0|X=A]\right)\\
&= \frac{2A}{1+A^2}Q(A),
\end{aligned}\tag{33}$$

where (in)-equalities follow from: a) using the bound in (32); b) using the fact that $\mathbb{E}[X|Y]\leq A$ and, therefore, $|A-\mathbb{E}[X|Y]|=(A-\mathbb{E}[X|Y])$; and c) using the bound $\mathbb{E}[X|Y=y]\leq A\cdot 1_{\{Y\geq 0\}}(y)$.

Since $g(x)$ is analytic, according to (33) there exists an interval around $x=A$ such that the derivative of $g(x)$ is strictly positive on that interval. Therefore, we can always find a δ_A (independent of the distribution on X) such that for all $x\in[A-\delta_A,A]$

$$\frac{dg(x)}{dx}>0.\tag{34}$$

Next, we use this fact to show that collapsing probabilities on the interval $[A,A-\delta_A]$ into mass points at A increases the MMSE. For any distribution F_X of X let

$$\bar{F}_X(x)=\begin{cases}F_X(-A+\delta_A) & -A\leq x\leq -A-\delta\\ F_X(x) & -A+\delta_A<x<A-\delta_A\\ \lim_{x\uparrow A-\delta_A}F_X(x) & A-\delta_A\leq x<A\\ 1 & x\geq A\end{cases}.$$

Note the construction of $\bar{F}_X(x)$ collapses the probabilities on intervals of length δ_A around $\pm A$ of a distribution F_X into mass points at $\pm A$. Moreover, observe that for any function $f(x)$ which is increasing on $x\in[-A,-A+\delta]$ and decreasing on $x\in[A-\delta_A,A]$ we have that

$$\mathbb{E}_{F_X}[f(X)]\leq\mathbb{E}_{\bar{F}_X}[f(X)].\tag{35}$$

Next, observe that

$$\begin{aligned}
\mathrm{mmse}(F_X,P_{Y|X}) &\overset{a)}{\leq} \mathbb{E}_{F_X}[(X-\mathbb{E}_{\bar{F}_X}[X|Y])^2]\\
&= \mathbb{E}_{F_X}\left[\mathbb{E}[(X-\mathbb{E}_{\bar{F}_X}[X|Y])^2|X]\right]\\
&\overset{b)}{\leq} \mathbb{E}_{\bar{F}_X}\left[\mathbb{E}[(X-\mathbb{E}_{F_X^\star}[X|Y])^2|X]\right]\\
&= \mathrm{mmse}(\bar{F}_X,P_{Y|X}),
\end{aligned}$$

where inequalities follow from: a) using the suboptimal estimator for F_X or the Pythagorean property in Theorem 1; and b) using the inequality in (35) by noting that $\mathbb{E}[(X-\mathbb{E}_{\bar{F}_X}[X|Y])^2|X=x]=g(x)$ which according to (34) is function for $x=[A,A-\delta_A]$.

This concludes the proof of the fact that the optimal input distribution must contain mass points at $\pm A$.

To show optimality of $X=\{\pm A\}$ for all $A\leq\bar{A}$ we use the necessary and sufficient condition for optimality given in (20). Using the fact that for $X=\{\pm A\}$ the conditional expectation is given by $\mathbb{E}[X|Y=y]=A\tanh(Ay)$ the condition in (20) can be further simplified to

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\int \mathrm{e}^{-\frac{(y-x)^2}{2}}(A^2\tanh^2(Ay)-2Ax\tanh(Ay))\\
&+\mathrm{e}^{-\frac{(y-A)^2}{2}}A^2\tanh(Ay)\mathrm{d}y+x^2-A^2\leq 0,\ \forall x\in[-A,A].
\end{aligned}\tag{36}$$

From (36) we see that by plugging $x=A$ (or $x=-A$) the necessary and sufficient condition for the optimality in (20b) is satisfied. Moreover, from (36), the largest A such that the second necessary and sufficient condition in (20a) is satisfied is given by $\bar{A}_B\approx 1.05647$.

■

For the case of $n\geq 1$ we have the following generalization of Proposition 3.

Definition 4. *A random vector* $\mathbf{X}$ *with a distribution* $P_{\mathbf{X}}$ *is said to be spherically symmetric if for every orthogonal matrix* $\mathbf{A}$ *we have that* $P_{\mathbf{X}}=P_{\mathbf{AX}}$.

Proposition 4. (Multivariate Gaussian.) *Let* $P_{\mathbf{Y}|\mathbf{X}}=\mathcal{N}(\mathbf{x},\mathbf{I})$ *and let*

$$\mathcal{C}(r)=\{\mathbf{x}\in\mathbb{R}^n:\|\mathbf{x}\|=r\}.\tag{37a}$$

Then for the optimization problem

$$\max_{P_{\mathbf{X}} \in \mathcal{F}_\infty(\mathcal{B}_{\mathbf{0}}(R))} \operatorname{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}), \tag{37b}$$

we have the following:

- *the optimal input distribution $P_{\mathbf{X}}^\star$ is unique and spherically symmetric. Moreover,*

$$\mathcal{E}(P_{\mathbf{X}}^\star) = \bigcup_{i=1}^{N} \mathcal{C}(r_i), \tag{38}$$

  where $N < \infty$ (finite) for some $\{r_i\}_1^N$;
- *$\mathcal{C}(R) \subseteq \mathcal{E}(P_{\mathbf{X}}^\star)$ for every $R > 0$;*
- *A uniform distribution over $\mathcal{C}(R)$ is optimal if and only if $R \leq \bar{R} = \Theta(\sqrt{n})$.*

Proof: The proof follows by mimicking the proof for the univariate case. The details are omitted and can be found in an extended version of this paper [9]. ∎

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 $\bar{R} \leq \sqrt{n}$.

Note that the result of Proposition 4 show that the optimal input distribution can be supported on the set $\mathcal{C}(R)$ which is a nowhere dense set of Lebesgue measure zero in $\mathbb{R}^n$. However, note that the set $\mathcal{C}(R)$ does have an uncountably infinite cardinality. Therefore, for $n > 1$ the conclusion in Proposition 2 is not superfluous and in general cannot be strengthen and discrete inputs are in general not optimal for $n > 1$.

Proposition 4 can be used to find the capacity of a MIMO channel given an amplitude constraint as follows.

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

$$\max_{\mathbf{X}:\mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}), \tag{39}$$

where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the optimizing input distribution is uniformly distributed on the set $\mathcal{C}(R) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = R\}$ (i.e., boundary of the ball) if $R \leq \sqrt{n}$.

Proof: Let $\mathbf{X}^\star \sim P_{\mathbf{X}}^\star$ be distributed on the boundary of the ball of radius R. First, observe that trivially we have that

$$\max_{\mathbf{X}:\mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}) \geq I(\mathbf{X}^\star; \mathbf{X}^\star + \mathbf{Z}).$$

Next, we show the upper bound. Using the I-MMSE relationship we have that

$$I(\mathbf{X}; \mathbf{X} + \mathbf{Z}) = \frac{1}{2} \int_0^1 \operatorname{mmse}(\mathbf{X}|\mathbf{Y}_\gamma) \mathrm{d}\gamma, \tag{40}$$

where $\mathbf{Y}_\gamma = \sqrt{\gamma}\mathbf{X} + \mathbf{Z}$. Next, let $\mathbf{W} = \sqrt{\gamma}\mathbf{X}$ and let $\mathbf{W}^\star = \sqrt{\gamma}\mathbf{X}^\star$ and observe that

$$\begin{aligned}
&\max_{\mathbf{X}:\mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} I(\mathbf{X}; \mathbf{X} + \mathbf{Z}) \\
&= \max_{\mathbf{X}:\mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} \frac{1}{2} \int_0^1 \operatorname{mmse}(\mathbf{X}|\mathbf{Y}_\gamma) \mathrm{d}\gamma \\
&\leq \frac{1}{2} \int_0^1 \max_{\mathbf{X}:\mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} \operatorname{mmse}(\mathbf{X}|\mathbf{Y}_\gamma) \mathrm{d}\gamma \\
&= \frac{1}{2} \int_0^1 \max_{\mathbf{X}:\mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\sqrt{\gamma}\mathbf{X} + \mathbf{Z}]\|^2\right] \mathrm{d}\gamma \\
&= \frac{1}{2} \int_0^1 \max_{\mathbf{X}:\mathbf{X} \in \mathcal{B}_{\mathbf{0}}(R)} \frac{1}{\gamma}\mathbb{E}\left[\|\sqrt{\gamma}\mathbf{X} - \mathbb{E}[\sqrt{\gamma}\mathbf{X}|\sqrt{\gamma}\mathbf{X} + \mathbf{Z}]\|^2\right] \mathrm{d}\gamma \\
&= \frac{1}{2} \int_0^1 \max_{\mathbf{W}:\in \mathcal{B}_{\mathbf{0}}(\sqrt{\gamma}R)} \frac{1}{\gamma}\mathbb{E}\left[\|\mathbf{W} - \mathbb{E}[\mathbf{W}|\mathbf{W} + \mathbf{Z}]\|^2\right] \mathrm{d}\gamma \\
&= \frac{1}{2} \int_0^1 \max_{\mathbf{W}:\in \mathcal{B}_{\mathbf{0}}(\sqrt{\gamma}R)} \frac{1}{\gamma}\mathbb{E}\left[\|\mathbf{W}^\star - \mathbb{E}[\mathbf{W}^\star|\mathbf{W}^\star + \mathbf{Z}]\|^2\right] \mathrm{d}\gamma \qquad (41) \\
&= \frac{1}{2} \int_0^1 \operatorname{mmse}(\mathbf{X}^\star|\mathbf{Y}_\gamma) \mathrm{d}\gamma \\
&= I(\mathbf{X}^*; \mathbf{X}^* + \mathbf{Z}),
\end{aligned}$$

where in (41) we have used that $\gamma \leq 1$ and the result in Proposition 4 that uniform distribution on the boundary of a ball of radius $\sqrt{\gamma}R \leq \sqrt{n}$. This concludes the proof. ∎

Note that Proposition 5 characterizes, previously unknown, capacity in the small amplitude regime (i.e., $R \leq \sqrt{n}$) in the massive MIMO case (i.e., the number of antennas going to infinity) [14].

E. Bounded Input: Poisson Noise Case

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

$$p_{Y|X}(y|x) = \frac{1}{y!} x^y \mathrm{e}^{-x},\ x \geq 0,\ y = 0, 1, ..., \tag{42}$$

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

$$\mathbb{E}[X|Y = y] = \frac{(y+1)p_Y(y+1; P_X)}{p_Y(y; P_X)},\ y = 0, 1, ..., \tag{43}$$

where $p_Y(y; P_X)$ is the marginal probability mass function (pmf) of Y induced by input distribution P_X. Next, we compute the MMSE for the binary input.

Lemma 4. *Let $X = \{0, A\}$ where $P_X[X = 0] = p_0$ and $P_{Y|X}$ be given as in* (42)*. Then,*

$$\mathbb{E}_{P_X}[X|Y = y] = \begin{cases} A\frac{(1-p_0)\mathrm{e}^{-A}}{p_0 + (1-p_0)\mathrm{e}^{-A}} & y = 0 \\ A & y > 0 \end{cases}, \tag{44a}$$

and

$$\operatorname{mmse}(X|Y) = \frac{A^2 \mathrm{e}^{-A} p_0 (1 - p_0)}{p_0 + (1 - p_0)\mathrm{e}^{-A}}. \tag{44b}$$

Proof: The transition probabilities of interest are given by

$$p_{Y|X}(y|0) = 1_{\{Y=0\}}(y),$$
$$p_{Y|X}(y|A) = \frac{1}{y!}A^y \mathrm{e}^{-A},$$

and the marginal of Y is given by

$$\begin{aligned} p_Y(y) &= p_0 p_{Y|X}(y|0) + (1-p_0)p_{Y|X}(y|A) \\ &= p_0 1_{\{Y=0\}}(y) + (1-p_0)\frac{1}{y!}A^y \mathrm{e}^{-A}. \end{aligned} \tag{45}$$

Therefore, by using (43) the estimator is given by

$$\begin{aligned} \mathbb{E}[X|Y=y] &= \frac{(y+1)p_Y(y+1)}{p_Y(y)} \\ &= \begin{cases} A\frac{(1-p_0)\mathrm{e}^{-A}}{p_0+(1-p_0)\mathrm{e}^{-A}} & y=0 \\ A & y>0 \end{cases}. \end{aligned} \tag{46}$$

To compute the MMSE observe that

$$\begin{aligned} \mathbb{E}[(X-\mathbb{E}[X|Y])^2] &= p_0 \mathbb{E}[(X-\mathbb{E}[X|Y])^2|X=0] \\ &\quad + (1-p_0)\mathbb{E}[(X-\mathbb{E}[X|Y])^2|X=A]. \end{aligned} \tag{47}$$

Next, we individually compute the terms in (47) by plugging in (46). The first term is given by

$$\begin{aligned} \mathbb{E}[(X-\mathbb{E}[X|Y])^2|X=0] &= \mathbb{E}[(\mathbb{E}[X|Y])^2|X=0] \\ &= \sum_{y=0}^{\infty}(\mathbb{E}[X|Y=y])^2 p_{Y|X}(y|0) \\ &= (\mathbb{E}[X|Y=0])^2 \\ &= A^2\left(\frac{(1-p_0)\mathrm{e}^{-A}}{p_0+(1-p_0)\mathrm{e}^{-A}}\right)^2, \end{aligned} \tag{48}$$

and the second term is given by

$$\begin{aligned} &\mathbb{E}[(X-\mathbb{E}[X|Y])^2|X=A] \\ &= \sum_{y=0}^{\infty}(A-\mathbb{E}[X|Y=y])^2 p_{Y|X}(y|A) \\ &= (A-\mathbb{E}[X|Y=0])^2 p_{Y|X}(0|A) \\ &= \left(A - A\frac{(1-p_0)\mathrm{e}^{-A}}{p_0+(1-p_0)\mathrm{e}^{-A}}\right)^2 \mathrm{e}^{-A} \\ &= A^2\left(\frac{p_0}{p_0+(1-p_0)\mathrm{e}^{-A}}\right)^2 \mathrm{e}^{-A}. \end{aligned} \tag{49}$$

Combining (48) and (49) we have that the MMSE is given by

$$\mathrm{mmse}(X|Y) = \frac{A^2\mathrm{e}^{-A}p_0(1-p_0)}{p_0+(1-p_0)\mathrm{e}^{-A}}.$$

This concludes the proof. ∎

Proposition 6. (Poisson Noise Case.) *Let $P_{Y|X}$ be as in* (42)*. Then for the optimization problem*

$$\max_{P_X \in \mathcal{F}_\infty([0,A])} \mathrm{mmse}(P_X, P_{Y|X}), \tag{50}$$

we have the following:

- *the optimal input distribution in* (50) *is discrete with finitely many points; and*
- *a two point distribution $\{0, A\}$ (i.e., $\mathcal{E}(P_X^\star) = \{0, A\}$) is optimal 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*

$$\mathrm{mmse}(P_X^\star, P_{Y|X}) = A^2\left(P_X^\star[X=0]\right)^2 = \frac{A^2}{\left(1+\mathrm{e}^{\frac{A}{2}}\right)^2}. \tag{51}$$

Proof: Let $p_Y(y; P_X)$ be the output pmf induced by the input distribution P_X

$$p_Y(y; P_X) = \int \frac{1}{y!}x^y \mathrm{e}^{-x} dP_X. \tag{52}$$

Note that the conditional expectation is given by Robins' formula as

$$\mathbb{E}_{P_X}[X|Y=y] = \frac{(y+1)p_Y(y+1; P_X)}{p_Y(y; P_X)}. \tag{53}$$

Moreover, the function $g(x)$ is given by

$$\begin{aligned} g(x) &= \mathbb{E}[(X-\mathbb{E}_{P_X}[X|Y])^2|X=x] \\ &= \sum_{y=0}^{\infty}\left(x-\mathbb{E}_{P_X}[X|Y=y]\right)^2 \frac{x^y \mathrm{e}^{-x}}{y!} \\ &= \mathrm{e}^{-x}\sum_{y=0}^{\infty}\left(x-\frac{(y+1)p_Y(y+1;P_X)}{p_Y(y;P_X)}\right)^2 \frac{x^y}{y!}. \end{aligned} \tag{54}$$

Clearly, the power series in (54) converges for all $x \in \mathbb{R}^+$ and $g(x)$ is an analytic function on $\mathbb{R}^+$ for any P_X. Also, evidently $g(x)$ is non-constant. Therefore, by Proposition 2 we have that the optimal input distribution is discrete with finitely many points.

Next we check whether distribution on $X = \{0, A\}$ with $P_X[X=0] = p_0$ is optimal by evaluating the necessary and sufficient conditions in (20). To that end, observe that with the estimator in (44a) $g(x)$ simplifies to

$$\begin{aligned} &g(x) \\ &= \mathbb{E}[(X-\mathbb{E}[X|Y])^2|X=x] \\ &= (x-\mathbb{E}[X|Y=0])^2\mathrm{e}^{-x} + \sum_{y=1}^{\infty}(x-A)^2\frac{1}{y!}x^y\mathrm{e}^{-x} \\ &= (x-\mathbb{E}[X|Y=0])^2\mathrm{e}^{-x} + (x-A)^2(1-\mathrm{e}^{-x}) \\ &= \left(x - A\frac{(1-p_0)\mathrm{e}^{-A}}{p_0+(1-p_0)\mathrm{e}^{-A}}\right)^2 \mathrm{e}^{-x} + (x-A)^2(1-\mathrm{e}^{-x}). \end{aligned} \tag{55}$$

First, the condition in (20b) is equivalent to

$$\text{eq. (44b)} = \text{eq. (48)} = \text{eq. (49)}, x \in \{0, A\}, \tag{56}$$

With some algebra (56) implies that the only possible value of p_0 is given by

$$p_0^2 = (1-p_0)^2 \mathrm{e}^{-A} \Leftrightarrow p_0 = \frac{1}{1+\mathrm{e}^{\frac{A}{2}}}. \tag{57}$$

With the choice of p_0 in (57) the MMSE in (44b) reduces to

$$\mathrm{mmse}(X|Y) = \frac{A^2}{\left(1+\mathrm{e}^{\frac{A}{2}}\right)^2}. \tag{58}$$

Second, the condition in (20a) requires that for all $x \in [0, A]$

$$\text{eq. (55)} \leq \text{eq. (58)}, \tag{59}$$

which can be further simplified to

$$\left(x - \frac{A}{1+\mathrm{e}^{\frac{A}{2}}}\right)^2 \mathrm{e}^{-x} + (x-A)^2(1-\mathrm{e}^{-x}) \leq \frac{A^2}{\left(1+\mathrm{e}^{\frac{A}{2}}\right)^2}. \tag{60}$$

for $x \in [0, A]$. It is not difficult to check that the condition in (60) fails if and only if the derivative of the function

$$h(x) \triangleq \left(x - \frac{A}{1+\mathrm{e}^{\frac{A}{2}}}\right)^2 \mathrm{e}^{-x} + (x-A)^2(1-\mathrm{e}^{-x}) - \frac{A^2}{\left(1+\mathrm{e}^{\frac{A}{2}}\right)^2}, \tag{61}$$

at zero becomes positive which occurs at values of A given by

$$2\mathrm{e}^{\frac{A}{2}}(A-1) + A\mathrm{e}^{A} - 2 = 0. \tag{62}$$

The solution to (62) is given by $\bar{A} \approx 0.9129$. Therefore, the binary input $\{0, A\}$ is optimal if and only if p_0 is the given in (57) and $A \leq \bar{A}$. This concludes the proof. ∎

F. Generalized Input Moment Constraints

In this section we seek to find

$$\sup_{P_{\mathbf{X}} \in \mathcal{F}_p(f;\alpha)} \mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) \tag{63a}$$

$$\text{where } \mathcal{F}(f;\alpha) = \left\{P_{\mathbf{X}} : \mathbb{E}_{P_{\mathbf{X}}}[f(\mathbf{X})] \leq \alpha\right\}. \tag{63b}$$

for some given $f : \mathbb{R}^n \to \mathbb{R}$ independent of $P_{\mathbf{X}}$ and given $p, \alpha \geq 0$. 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 Markov inequality and Prokhorov theorems implies that $\mathcal{F}(f;\alpha)$ is a sequentially compact set. An example of $f(\cdot)$ that satisfies such a condition is $f(\mathbf{X}) = \|\mathbf{X}\|^r$ for any $r > 0$.

Theorem 7. *Let the MMSE in the optimization problem in* (63) *be an upper semicontinuous function. Then, the supremum in* (63a) *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 $\mathbf{x} \in \mathbb{R}^n$*

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] - \lambda\left(f(\mathbf{x}) - \alpha\right) \leq \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}); \text{ and} \tag{64a}$$

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

$$\mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] - \lambda\left(f(\mathbf{x}) - \alpha\right) = \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}). \tag{64b}$$

Proof: First, in light of the facts that $\mathcal{F}(f;\alpha)$ is a convex and sequentially compact set, and that the MMSE is an upper semicontinuous function by property 1) of Theorem 4, we have that the supremum in (63a) is attained by some distribution $P_{\mathbf{X}}^\star$. Since the MMSE is a concave function and the constraint in (63b) is linear, using the KKT conditions in Theorem 4 we have that the constrained optimization in (63) is equivalent to

$$\sup_{P_{\mathbf{X}} \in \mathcal{F}} \left(\mathrm{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) - \lambda\left(\mathbb{E}_{P_{\mathbf{X}}}[f(\mathbf{X})] - \alpha\right)\right). \tag{65}$$

Moreover, from (17) we have that $\lambda > 0$ this follows since the constraint in (63b) is tight and $\mathbb{E}_{P_{\mathbf{X}}^\star}[f(\mathbf{X})] = \alpha$.

Next, since the difference of concave and linear functions is concave we have that the function in (65) is concave. Therefore, applying property 3) of Theorem 4 to (65) we have that the input distribution $P_{\mathbf{X}}^\star$ is optimal if and only if

$$\mathrm{mmse}_{Q_{\mathbf{X}}}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) - \mathrm{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) - \lambda\left(\mathbb{E}_{Q_{\mathbf{X}}}[f(\mathbf{X})] - \alpha\right) \leq 0, \tag{66}$$

where we have used that the Gâteaux derivative of $\mathbb{E}_{P_{\mathbf{X}}}[f(\mathbf{X})]$ is given by

$$\Delta_{Q_{\mathbf{X}}} \mathbb{E}_{P_{\mathbf{X}}}[f(\mathbf{X})] = \mathbb{E}_{Q_{\mathbf{X}}}[f(\mathbf{X})] - \mathbb{E}_{P_{\mathbf{X}}}[f(\mathbf{X})],$$

and that for the optimal input distribution $\mathbb{E}_{P_{\mathbf{X}}^\star}[f(\mathbf{X})] = \alpha$.

The proof is concluded by using the approach in Proposition 1 to show that the condition (66) can be equivalently represented as in (64). ∎

Proposition 7. *Let $P_{\mathbf{Y}|\mathbf{X}} = \mathcal{N}(\mathbf{x}, \mathbf{I})$. Then for the optimization problem in* (63) *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 optimizing 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 optimizing distribution is unbounded (i.e., there is no $R \geq 0$ such that $\mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathcal{B}_{\mathbf{0}}(R)$).*

Proof: We first work by assuming that $f(\mathbf{x}) = \omega\left(\|\mathbf{x}\|^2\right)$. Next, towards a contradiction assume that the maximizing distribution in (63) has an unbounded constraint. In other words, there exists no $R > 0$ such that $\mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathcal{B}_0(R)$.

By using the KKT condition in (64b) we have that for all $\mathbf{x} \in \mathcal{E}(P_{\mathbf{X}}^\star)$

$$\begin{aligned}
&\operatorname{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) + \lambda\left(f(\mathbf{x}) - \alpha\right) \\
&= \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \\
&\overset{a)}{=} \mathbb{E}\left[\|\mathbf{Z} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{Z}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \\
&\overset{b)}{\leq} 2(\mathbb{E}\left[\|\mathbf{Z}\|^2\right] + \mathbb{E}\left[\mathbb{E}_{P_{\mathbf{X}}^\star}[\|\mathbf{Z}\|^2|\mathbf{Y}]|\mathbf{X} = \mathbf{x}\right]) \\
&\overset{c)}{\leq} 2n + 2\mathbb{E}\left[c\left(\|\mathbf{Y}\|^2 + \mathbb{E}_{P_{\mathbf{X}}^\star}[\|\mathbf{X}\|^2]\right)|\mathbf{X} = \mathbf{x}\right] \\
&= 2n + 2c(n + \|\mathbf{x}\|^2 + \mathbb{E}_{P_{\mathbf{X}}^\star}[\|\mathbf{X}\|^2]), \qquad (67)
\end{aligned}$$

where the (in)-equalities follow from: a) using $\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}] = -(\mathbf{Z} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{Z}|\mathbf{Y}])$; b) using the bound $\|\mathbf{a} - \mathbf{b}\|^2 \leq 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$ and modulus inequality $\|\mathbb{E}[\mathbf{U}]\| \leq \mathbb{E}[\|\mathbf{U}\|]$ for any $\mathbf{U}$; and c) using the bound $\mathbb{E}_{P_{\mathbf{X}}^\star}[\|\mathbf{Z}\|^2|\mathbf{Y} = \mathbf{y}] \leq c(\|\mathbf{y}\|^2 + \mathbb{E}_{P_{\mathbf{X}}^\star}[\|\mathbf{X}\|^2])$ for some fixed constant $c > 0$ [10, Lemma 4].

Observe that the inequality in (67) implies that for all $\mathbf{x} \in \mathcal{E}(P_{\mathbf{X}}^\star)$

$$f(\mathbf{x}) \leq a_1\|\mathbf{x}\|^2 + a_2, \qquad (68)$$

for some fixed constant $a_1 > 0$ and a_2. Since, we are assuming that $\mathcal{E}(P_{\mathbf{X}}^\star)$ is unbounded that means that there exists a sequence $\{\mathbf{x}_n\}_{n=1}^\infty \subseteq \mathcal{E}(P_{\mathbf{X}}^\star)$ such that $\|\mathbf{x}_n\| \to \infty$. However, the existence of such a sequence together with the inequality in (68) contradicts our assumption $f(\mathbf{x}) = \omega\left(\|\mathbf{x}\|^2\right)$. Therefore, $\mathcal{E}(P_{\mathbf{X}}^\star)$ must be bounded, and this concludes the proof for the case of $f(\mathbf{x}) = \omega\left(\|\mathbf{x}\|^2\right)$.

The case of $f(\mathbf{x}) = \|\mathbf{x}\|^2$ (i.e., power constraint) is well known in the literature for example see [15].

Finally, we look at the case of $f(\mathbf{x}) = o(\|\mathbf{x}\|^2)$. Towards a contradiction we assume that $\mathcal{E}(P_{\mathbf{X}}^\star) \subseteq \mathcal{B}_0(R)$ for some $R > 0$. Using the KKT condition in (64a) we have that for all $\mathbf{x} \in \mathbb{R}^n \setminus \mathcal{B}_0(R)$

$$\begin{aligned}
&\operatorname{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) + \lambda\left(f(\mathbf{x}) - \alpha\right) \\
&\geq \mathbb{E}\left[\|\mathbf{X} - \mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \\
&= \|\mathbf{x}\|^2 - 2\mathbf{x}^T\mathbb{E}\left[\mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]|\mathbf{X} = \mathbf{x}\right] \\
&+ \mathbb{E}\left[\|\mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2 | \mathbf{X} = \mathbf{x}\right] \\
&\overset{a)}{\geq} \|\mathbf{x}\|^2 - 2\|\mathbf{x}\|\left\|\mathbb{E}\left[\mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]|\mathbf{X} = \mathbf{x}\right]\right\| \\
&\overset{b)}{\geq} \|\mathbf{x}\|^2 - 2\|\mathbf{x}\|R, \qquad (69)
\end{aligned}$$

where the inequalities follow from: a) using the bound $\mathbb{E}\left[\|\mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]\|^2|\mathbf{X} = \mathbf{x}\right] > 0$ and applying Cauchy-Swartz inequality for the inner product; and b) using the fact that $\|\mathbf{X}^\star\| \leq R$ and therefore $\left\|\mathbb{E}\left[\mathbb{E}_{P_{\mathbf{X}}^\star}[\mathbf{X}|\mathbf{Y}]|\mathbf{X} = \mathbf{x}\right]\right\| \leq \mathbb{E}\left[\mathbb{E}_{P_{\mathbf{X}}^\star}[\|\mathbf{X}\|\,|\mathbf{Y}]|\mathbf{X} = \mathbf{x}\right] \leq R$. Clearly the condition in (69) show $f(\mathbf{x}) = \Omega(\|\mathbf{x}\|^2)$. However, this contradicts our assumption that $f(\mathbf{x}) = o(\|\mathbf{x}\|^2)$. Therefore, $\mathcal{E}(P_{\mathbf{X}}^\star)$ must be unbounded. This concludes the proof. ∎

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.

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

V. Single Crossing Point Property

$$\max_{P_{\mathbf{X}} \in \mathcal{M}_2} \left(\operatorname{mmse}(P_{\mathbf{X}}, P_{\mathbf{Y}|\mathbf{X}}) - \lambda \operatorname{mmse}(P_{\mathbf{X}}, Q_{\mathbf{Y}|\mathbf{X}})\right). \qquad (70)$$

where

$$\mathcal{M}_2 \triangleq \{P_{\mathbf{X}} : \mathbf{K}_{\mathbf{X}} \preceq \mathbf{S}\} \qquad (71)$$

Then necessary condition for optimality is given by

$$\begin{aligned}
&\operatorname{mmse}_{F_{\mathbf{X}}}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) - \operatorname{mmse}(P_{\mathbf{X}}^\star, P_{\mathbf{Y}|\mathbf{X}}) \\
&- \lambda(\operatorname{mmse}_{F_{\mathbf{X}}}(P_{\mathbf{X}}^\star, Q_{\mathbf{Y}|\mathbf{X}}) - \operatorname{mmse}(P_{\mathbf{X}}^\star, Q_{\mathbf{Y}|\mathbf{X}})) \leq 0,\ \forall F_{\mathbf{X}} \in \mathcal{M}_2 \qquad (72)
\end{aligned}$$

VI. Conclusion

In this work we looked at the structure of the support of least favorable prior distributions. We demonstrated that, under some mild conditions, the support of the least favorable distributions must be a nowhere dense set of Lebesgue measure zero. Our results also produce necessary and sufficient conditions for the 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 look at the optimization problem where for $\lambda \geq 0$ 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 [15] and discussed in great detail in [16] and [17], to the vector cases.

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, April 2005.

[3] M. Raginsky, "On the information capacity of gaussian channels under small peak power constraints," in *46th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 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). (2018) "On the Structure of the Least Favorable Prior Distributions". [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, March 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] R. M. Dudley, *Real analysis and probability*. Cambridge University Press, 2002, vol. 74.
[14] 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*. IEEE, 2017, pp. 1058–1062.
[15] 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, April 2011.
[16] 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, Feb 2013.
[17] 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.