When Are Discrete Channel Inputs Optimal? – Optimization Techniques and Some New Results

Alex Dytso*, Mario Goldenbaum*, H. Vincent Poor*, and Shlomo Shamai (Shitz)†

*Department of Electrical Engineering, Princeton University
†Department of Electrical Engineering, Technion – Israel Institute of Technology

***Abstract*—Discrete channel inputs have been shown to maximize mutual information under various settings. This paper offers a brief review of corresponding optimization methods. Specifically, two techniques are considered; the first is rooted in the theory of linear optimization whereas the second is based on ideas from convex optimization.**

I. Introduction

In this work, we are interested in the optimization problem

$$\max_{X:X\sim P_X\in\mathcal{M}} I(X;Y), \tag{1}$$

where $I(X;Y)$ is the mutual information between random vectors X and Y distributed according to some joint distribution P_{XY}. Their respective marginal distributions are denoted by P_X (input distribution) and P_Y (output distribution), respectively, whereas $\mathcal{M}$ is some compact and convex set of input distributions.[1] We are specifically interested in an exact characterization of the optimizing distribution $P_X^\star$ (not necessarily unique). In cases where an exact optimizer cannot be found we determine properties of an optimal input distribution in order to reduce the dimensionality of the problem. For example, in many cases it turns out that an optimizing distribution is discrete with finitely many mass points, which reduces the infinite dimensional problem (1) to a finite dimensional one allowing to apply numerical methods for computing $P_X^\star$.

Towards this end, we survey two popular approaches to characterize the maximum in (1). The first approach assumes that the output space is of finite cardinality whereas the input space can be of arbitrary size. The technique is based on the theory of linear optimization. We also introduce several new techniques to this approach. For example, the new tools not only allow to work with Euclidean input spaces but also with inputs supported on separable metric spaces.

The second approach makes no assumption on the cardinality of the chanenl output space and is based on convex optimization techniques. This approach has usually been applied to channels with scalar inputs. We introduce several new tools in order to extend the results to the vector case. Moreover, in cases in which the input distribution is discrete with finitely many mass points, we show how to obtain an estimate on the number of points.

Due to space limitations, we do not focus on the algorithmic aspects of finding an optimal input distribution. The literature on this subject is vast and the interested reader is referred to [1]–[3] and references therein. We also do not focus on the popular subject of approximating the capacity of a channel with discrete inputs (see [4]–[6] and references therein).

The paper is organized as follows. Section II introduces a preliminary set of mathematical tools needed in our analysis such as tools from probability theory, convex analysis, and mathematical optimization. Section III then presents the first optimization technique that focuses on channel output spaces of finite cardinality. Under very general assumptions it is shown that the optimizing input distribution must be discrete with a finite number of mass points. Moreover, it is shown that the number of mass points growth linearly with the cardinality of the output space and linearly with the number of constraints on the input. The second optimization technique is presented in Section IV and it is shown under very mild conditions that in the support of the optimal input distribution is generally a nowhere dense set of Lebesgue measure zero. Moreover, in cases in which the optimal input distribution is discrete with finitely many mass points it is shown how the number of points can be determined. Finally, Section V concludes the paper.

Notation: deterministic quantities (i.e, scalars and vectors) are denoted by lower case letters and random objects by capital letters; $\mathbb{R}$ denotes the affinely extended real number system; the closed ball in $\mathbb{R}^n$ of radius R centered at x is denoted as $\mathcal{B}_x(R) \coloneqq \{y \in \mathbb{R}^n : \|y-x\| \leq R\}$, where $\|\cdot\|$ is the Euclidean norm; the Dirac measure centered on a fixed point x is denoted as δ_x; the mutual information between X and Y distributed according to P_{XY} is also denoted as $I(P_X, P_{Y|X})$.

II. Some Elementary Definitions and Results

In this section, we recap some elementary definitions and results from probability theory and mathematical optimization that will be very useful for our purposes.

For a random vector $X \in \mathbb{R}^n$ and every measurable set $\mathcal{A} \subset \mathbb{R}^n$ we denote the probability measure of X as

$$P_X(\mathcal{A}) = \mathbb{P}[X \in \mathcal{A}].$$

If it is clear from the context we sometimes write P instead of P_X. The space of all probability measures defined on a sample space $\Omega \subseteq \mathbb{R}^n$ is denoted as $\mathcal{P}(\Omega)$.

This work was supported in part by the U. S. National Science Foundation under Grants CCF-1420575 and CNS-1456793, by the German Research Foundation under Grant GO 2669/1-1, and by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630.

[1]Note that $\mathcal{M}$ is very general in the sense that it accounts for any type of constraints on the channel inputs (e.g., amplitude constraints, average power constraints).

978-1-5386-0579-0/18/$31.00 ©2018 IEEE

A. Weak Convergence and Weak Continuity

It is well known that there exist several different notions of the convergence of a sequence of probability measures. One of them is weak convergence, which provides a given space of probability measures with a topology.

Definition 1. A sequence of probability measures $\{P_n\}_{n\in\mathbb{N}}$ is said to *converge weakly* to the probability measure P if

$$\lim_{n\to\infty} \mathbb{E}_{P_n}\left[\phi(X)\right] \to \mathbb{E}_P\left[\phi(X)\right], \tag{2}$$

for all bounded and continuous functions ϕ.

Another main ingredient of our considerations are linear functionals. The following theorem gives a necessary and sufficient condition for a linear functional to be weakly continuous [7, Lemma 2.1].

Theorem 1. (Weak Continuity of Linear Functionals) *A linear functional $L : \mathcal{P} \to \mathbb{R}$ is* weakly continuous *on $\mathcal{P}$ if and only if it can be represented as*

$$L(P) = \mathbb{E}_P[\phi(X)],$$

for some bounded and continuous function ϕ.

Definition 2. A given space of probability measures $\mathcal{P}(\Omega)$ is said to be *compact* if every infinite sequence, $\{P_n\}_{n\in\mathbb{N}}$, in $\mathcal{P}(\Omega)$ has a weakly convergent subsequence.[2]

B. Elementary Optimization Theorems

The extreme value theorem for real-valued continuous functions over compact intervals is one of the most celebrated results of calculus. The following theorem is a generalization to compact topological spaces [9, Sec. 2.13].

Theorem 2. (Extreme Value Theorem) *For any compact topological space $\mathcal{P}$ and any upper semicontinuous functional $f : \mathcal{P} \to \mathbb{R}$*

$$\sup_{P\in\mathcal{P}} f(P) = \max_{P\in\mathcal{P}} f(P).$$

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

Definition 3. An *extreme point* of any convex set $\mathcal{S}$ is a point $x \in \mathcal{S}$ that cannot be represented as $x = (1-t)y + tz$ with $y, z \in \mathcal{S}$ and some $t \in (0, 1)$. We denote the set of all extreme points of $\mathcal{S}$ as $\mathrm{ex}\mathcal{S}$.

As in this paper we are interested in the extreme points of certain sets of probability measures, the following theorem will be of particular importance [10, Th. 2.1].

Theorem 3. (Extreme Points of Moment Sets) *Let $(\Omega, \sigma(\Omega))$ be a measurable space and let $\mathcal{P}_{\mathsf{reg}}(\Omega)$ be the set of all regular probability measures over the sample space Ω.*[3] *Fix measurable functions $f_1, \ldots, f_k$ as well as real numbers $c_1, \ldots, c_k$ and consider the set*

$$\mathcal{H}_k \coloneqq \{P \in \mathcal{P}_{\mathsf{reg}}(\Omega) : \mathbb{E}_P[f_i(X)] \le c_i,\, 1 \le i \le k\}; \tag{3}$$

that is, the set of regular probability measures with $k \in \mathbb{N}$ bounded moments (called moment set). Then,

1) $\mathcal{H}_k$ is convex and the extreme points of $\mathcal{H}_k$ are given by

$$\mathrm{ex}\mathcal{H}_k \subseteq \overline{\mathrm{ex}\mathcal{H}}_k, \tag{4}$$

where

$$\overline{\mathrm{ex}\mathcal{H}}_k \coloneqq \Big\{P \in \mathcal{H}_k : P = \sum_{i=1}^{m} \alpha_i \delta_{x_i}, x_i \in \Omega, \alpha_i \in [0,1], \sum_{i=1}^{m} \alpha_i = 1, 1 \le m \le k+1, \text{the vectors } [f_1(x_i), \ldots, f_k(x_i), 1], 1 \le i \le m, \text{are linearly independent}\Big\};$$

2) If the moment conditions in (3) *are fulfilled with equality, then* (4) *holds with equality.*

The following result, taken from [10, Th. 3.2], states that when maximizing a linear functional over a moment set it is sufficient to focus on its extreme points.

Theorem 4. (Linear Programming) *Let $L : \mathcal{P} \to \mathbb{R}$ be a linear functional. Then,*

$$\sup_{P\in\mathcal{H}_k} L(P) = \sup_{P\in\mathrm{ex}\mathcal{H}_k} L(P).$$

Note that Theorem 4 only requires L to be linear and not necessarily continuous.

Definition 4. A convex set $\mathcal{S}$ of a vector space $\mathcal{V}$ is called *linearly closed (linearly bounded)* if every straight line intersects with $\mathcal{S}$ on a closed (bounded) subset of that line.

With this definition in hand, we close the section with a theorem proven by Dubin [11].

Theorem 5. (Dubins Theorem) *Let $f : \mathcal{V} \to \mathbb{R}$ be a linear functional over a vector space $\mathcal{V}$ and let*

$$\mathcal{L} = \{v \in \mathcal{V} : f(v) = c\},$$

for some constant $c \in \mathbb{R}$, be a hyperplane formed by f. Moreover, let $\mathcal{I}$ be the intersection of a linearly closed and linearly bounded convex set $\mathcal{K} \subset \mathcal{V}$ with n hyperplanes. Then, every extreme point of $\mathcal{I}$ is a convex combination of at most $n+1$ extreme points of $\mathcal{K}$.

A remarkable property of Theorem 5 is that it also holds for the infinite dimensional case.

III. Channel Outputs of Finite Cardinality

In this section, we consider the optimization problem (1) for channels whose outputs are of bounded cardinality. Towards this end, we are going to generalize an elegant approach proposed by Witsenhausen in [12]. The approach relies on techniques from linear optimization and convex analysis.

[2] More precisely, the space is *sequentially compact*. By means of Prohorov's metric, however, the space of probability measures equipped with the topology of weak convergence becomes a metric space [7, Th. 3.8], in which case the notions compact and sequentially compact are equivalent [8, Th. 2.3.1].

[3] Recall that a probability measure is *regular* if any element of the σ-algebra $\sigma(\Omega)$ can be approximated from below by compact measurable sets and from above by open measurable sets.

As the main result of this section, the following theorem characterizes under very mild conditions the structure of the optimal input distribution.

Theorem 6. (Optimal Input Distribution Under Finite Output Cardinality) *Let $N \in \mathbb{N}$ be finite and assume the following:*

- *$\mathcal{H}_k$ as defined in (3) is compact;*
- *$|\mathsf{supp}(Y)| \leq N$ (i.e., the support of Y is finite);*
- *$P_{Y|X}(y|x)$ is continuous in x for every y.*

Then, there exists a distribution $P_X^\star \in \mathcal{H}_k$ so that

$$\sup_{P_X \in \mathcal{H}_k} I(P_X, P_{Y|X}) = I(P_X^\star, P_{Y|X}).$$

Moreover, $P_X^\star$ is discrete with at most $N(k+1)$ mass points (possibly containing points at $\pm\infty$) and we have the following:

- *If $\Omega \subset \mathbb{R}^n$ is compact and $f_1, \ldots, f_k$ are bounded and continuous on Ω, then $P_X^\star$ has at most $N+k$ finite mass points;*
- *If $f_1, \ldots, f_k$ are bounded and continuous on $\Omega = \mathbb{R}^n$ and are such that for every P_X with a finite number of mass points $\mathbb{E}_{P_X}[f_i(X)] < \infty$ implies $P_X(+\infty) = P_X(-\infty) = 0$, then $P_X^\star$ has at most $N+k$ finite mass points.*

Proof: First of all, since the moment set $\mathcal{H}_k$ is assumed to be compact and $P_X \mapsto I(X;Y)$ is concave, by Theorem 2 we have that the supremum is attained by some $P_X^\star \in \mathcal{H}_k$.

Next, we show that $P_X^\star$ is discrete and provide a bound on the number of probability masses. The assumption that Y is discrete allows us to write

$$I(P_X, P_{Y|X}) = H(Y) - H(Y|X). \tag{5}$$

Recall that the entropy of Y is given by

$$H(Y) = -\sum_{i=1}^{N} p_i \log(p_i),$$

where the elements of the vector of output probabilities $p_Y \coloneqq [p_1, \ldots, p_N]$ are of the form

$$p_i = \int P_{Y|X}(y_i|x)\, \mathrm{d}P_X,$$

so we can write p_Y as a linear transformation

$$p_Y = \langle P_{Y|X}, P_X \rangle.$$

Now, let $p_Y^\star \coloneqq [p_1^\star, \ldots, p_N^\star]$ denote the output vector induced by an optimal input distribution $P_X^\star$ and note that the conditional entropy is given by

$$H(Y|X) = -\int \sum_{i=1}^{N} P_{Y|X}(y_i|x) \log\big(P_{Y|X}(y_i|x)\big) \mathrm{d}P_X.$$

As $P_X \mapsto H(Y)$ is concave and $P_X \mapsto H(Y|X)$ linear, it follows that $I(P_X, P_{Y|X})$ is the difference between a concave and a linear functional.

Let the set of distributions that induce the optimal output distribution defined as

$$\mathcal{P}^\star \coloneqq \left\{ P_X \in \mathcal{H}_k : p_Y^\star = \langle P_{Y|X}, P_X \rangle \right\}. \tag{6}$$

Observe that $\mathcal{P}^\star$ is an intersection of $\mathcal{H}_k$ with the $N-1$ hyperplanes

$$\mathcal{L}_i \coloneqq \left\{ P_X : \int P_{Y|X}(y_i|x)\, \mathrm{d}P_X = p_i^\star \right\},$$

$i = 1, \ldots, N-1$. Note that we only consider $N-1$ hyperplanes instead of N. This is because in the space of probability distributions everything adds up to one so that $\mathcal{L}_N$ is redundant. Note also that each $\mathcal{L}_i$ is a closed set, which follows from Theorem 1 because $P_{Y|X}(y_i|x)$ is continuous and bounded in x.

As the intersection of a compact with a closed set is compact, we have that $\mathcal{P}^\star$ is compact. This implies that

$$\max_{P_X \in \mathcal{H}_k} I(P_X, P_{Y|X}) = \max_{P_X \in \mathcal{P}^\star} I(P_X, P_{Y|X}).$$

Moreover, it follows from (5) that

$$\begin{aligned} \max_{P_X \in \mathcal{P}^\star} I(P_X, P_{Y|X}) &= \max_{P_X \in \mathcal{P}^\star} \big(H(Y) - H(Y|X)\big) \\ &= H(Y^\star) + \max_{P_X \in \mathcal{P}^\star} \big(-H(Y^\star|X)\big), \end{aligned}$$

where the last step follows from the fact that all distributions in $\mathcal{P}^\star$ induce the same $H(Y)$.

Now, as $P_X \mapsto H(Y|X)$ is linear, by means of Theorem 4 we conclude that

$$\max_{P_X \in \mathcal{P}^\star} \big(-H(Y|X)\big) = \max_{P_X \in \mathrm{ex}\mathcal{P}^\star} \big(-H(Y|X)\big). \tag{7}$$

In the following, let $P_X^\star \in \mathrm{ex}\mathcal{P}^\star$ be a distribution that maximizes (7). Due to Theorem 3, we have that any $P_X \in \mathrm{ex}\mathcal{P}^\star$, and in particular $P_X^\star$, can be represented as a convex combination of at most $(N-1)+1 = N$ extreme points of $\mathcal{H}_k$. Due to Theorem 5, however, the extreme points are given by discrete distributions with at most $k+1$ points from which we conclude that $P_X^\star$ consists of at most $N(k+1)$ mass points.

Now that we know the maximizing input distribution is discrete with at most $N(k+1)$ mass points, we are able to slightly refine the number under various assumptions. Due to the lack of space, the corresponding proofs are deferred to the extended version of this paper. ∎

Remark 1. The original proof by Witsenhausen [12] was concerned with the scalar case only where $\Omega = [-a, a]$ for some $a > 0$. All the subsequent extension of this result relied on reducing the optimization problems to the case of bounded support. The key novelty in the proof of Theorem 6 is the application of Theorem 3, which does not require to show or assume Ω is bounded. In fact, Theorem 3 does not make use of the spaces underlying X and Y and holds if Ω is a well behaving set of some vectors space.

Remark 2. If the functions $f_1, \ldots, f_k$ in the definition of $\mathcal{H}_k$ prevent the occurrence of mass points at $\pm\infty$, then the bound on the number of points can be reduced to $N+k$. An example of such a function is $f(x) = |x|^r, r > 0$, which naturally forces probability measures with a finite number of mass points to have mass points at $\pm\infty$ with zero probability.

Corollary 1. *Let U be arbitrary but independent of X. Then, Theorem 6 is valid for the optimization problem*

$$\sup_{P_X \in \mathcal{H}_k} I(X;Y|U).$$

We close this section with a historical note. In the context of discrete memoryless channels, Gallager has shown in [13] that the cardinality of the input should not exceed the cardinality of the output. However, Gallager's result does not apply to the setting of Theorem 6 in which the input space Ω may not be discrete a priori. It also does not hold for general input constraints. The approach in this section has also been applied for point-to-point channels with output quantization [14], [15].

IV. Convex Optimization Approach

In this section, we follow the convex optimization method taken by Smith in [16]. Unlike Section III, however, we do not make any assumption on the channel output alphabets. In order to determine some properties of the optimal input distributions, we introduce tools with which we are able to obtain results for the multivariate case, which is in contrast to the majority of related works considering the scalar case only.

Whereas in the previous section we were able to obtain relatively tight bounds on the number of mass points an optimal distribution must have, to the best of our knowledge the approach of Smith has never been used to obtain similar bounds. In Section IV-B, we therefore introduce a method that can be used to bound the cardinality of the support of the optimal input distribution provided it is discrete.

A. Necessary and Sufficient Conditions for Optimality

The key tool of this section will the notion of weak or directional derivative over the space of probability distributions.

Definition 5. Let $\mathcal{P}$ be a convex topological space. For any two distributions $P \in \mathcal{P}$ and $Q \in \mathcal{P}$ we define the *Gâteaux derivative* of $f : \mathcal{P} \to \mathbb{R}$ at P in the direction of Q as

$$\Delta_Q f(P) = \lim_{\varepsilon \to 0} \frac{f\big((1-\varepsilon)P + \varepsilon Q\big) - f(P)}{\varepsilon}.$$

We will use the Gâteaux derivative together with the following optimization theorems.

Theorem 7. *Let $\mathcal{P}$ be a convex topological space and let $f : \mathcal{P} \to \mathbb{R}$ have a Gâteaux derivative $\Delta_Q f(P)$ for every $P, Q \in \mathcal{P}$. Suppose that $P^\star \in \mathcal{P}$ is a maximizer of f, then*

$$\forall Q \in \mathcal{P} : \Delta_Q f(P^\star) \leq 0. \tag{8}$$

If in addition f is concave, (8) *is also sufficient.*

Theorem 8. (Karush-Kuhn-Tucker Conditions) *Let $\mathcal{P}$ be a convex topological space, $f : \mathcal{P} \to \mathbb{R}$ a concave functional, and $g : \mathcal{P} \to \mathbb{R}$ a convex functional. Assume there exists a point $P \in \mathcal{P}$ such that $g(F) < 0$. Furthermore, let*

$$\mu \coloneqq \sup_{P \in \mathcal{P}, g(P) \leq 0} f(P). \tag{9}$$

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

$$\mu = \sup_{P \in \mathcal{P}} \big(f(P) - \lambda g(P)\big). \tag{10}$$

If the supremum in (9) *is attained by some P_0, then P_0 also attains the supremum in* (10) *with $\lambda g(P_0) = 0$.*

Unlike its counterparts in finite dimensions, the Gâteaux derivative may exist without the functional being continuous. For example, the Gâteaux derivative of a linear functional is of the form

$$\Delta_Q \mathbb{E}_P[f(X)] = \mathbb{E}_Q[f(X)] - \mathbb{E}_P[f(X)],$$

which exists as long as $\mathbb{E}_P[f(X)]$ and $\mathbb{E}_Q[f(X)]$ are finite. However, $P \mapsto \mathbb{E}_P[f(X)]$ is continuous if and only if f is bounded and continuous (see Theorem 1).

Another assumption that we make in this section is that the Gâteaux derivative of the mutual information exists.

Assumption 1. Suppose the Gâteaux derivative of the mutual information $P_X \mapsto I(P_X, P_{Y|X})$ exists for all $P_X, Q_X \in \mathcal{M}$ and that it is given by

$$\Delta_{Q_X} I(P_X, P_{Y|X}) = I_{Q_X}(P_X, P_{Y|X}) - I(P_X, P_{Y|X}), \tag{11}$$

where

$$I_{Q_X}(P_X|P_{Y|X}) \coloneqq \mathbb{E}_{Q_X}\left[\log\left(\frac{P_{Y|X}(Y|X)}{P_Y(Y;P_X)}\right)\right]$$

and $P_Y(Y;P_X)$ denoting the channel output distribution induced by the input distribution P_X.

We do not formally prove that the Gâteaux derivative is of the form (11). However, for all the known cases of interest it is given by this expression.

Next, we provide a necessary and sufficient condition for the optimality of an input distribution P_X.

Theorem 9. (Necessary and Sufficient Optimality Condition) *Let $\mathcal{M}$ be a convex and compact set of channel input distributions. Then, $P_X^\star \in \mathcal{M}$ maximizes* (1) *if and only if*

$$\forall Q_X \in \mathcal{M} : I_{Q_X}(P_X^\star, P_{Y|X}) \leq I(P_X^\star, P_{Y|X}).$$

Proof: The proof is based on Theorem 7, the Gâteaux derivative (11) and the concavity of mutual information. ■

B. Structure of the Support

In many cases of interest, the set $\mathcal{M}$ is equal to $\mathcal{H}_k$ with functions $f_1, \ldots, f_k$ chosen such that $\mathcal{H}_k$ is compact. In other words, $\mathcal{M}$ is a moment set, a set of distributions that are of compact support, or a combination of both. For ease of presentation, we focus on the case $\mathcal{M} = \mathcal{H}_1$ only. The results, however, are extendible to the cases $\mathcal{H}_k$, $k \in \mathbb{N}$.

In order to study the support of the optimal input distribution we will need the following definition.

Definition 6. A point $x \in \mathbb{R}^n$ is said to be *a point of increase* of a distribution P_X, if for any open subset $\mathcal{O} \subset \mathbb{R}^n$ containing x, $P_X(\mathcal{O}) > 0$. We denote the set of points of increase of P_X as $\mathcal{E}(P_X) \subseteq \mathbb{R}^n$.

Observe that $P_X\big(\mathcal{E}(P_X)\big) = 1$. In fact, $\mathcal{E}(P_X)$ is the smallest closed subset of $\mathbb{R}^n$ whose probability is 1.

Theorem 10. (Sufficient and Necessary Condition) *Let* $i(x;P_X,P_{Y|X}):\mathbb{R}^n\to\mathbb{R}$ *be defined as*

$$i(x;P_X,P_{Y|X}) \coloneqq \mathbb{E}\left[\log\left(\frac{P_{Y|X}(Y|X)}{P_Y(Y;P_X)}\right)\middle|X=x\right].$$

Then, $P_X^\star$ *is an optimizer if and only if there exists a* $\lambda\geq 0$ *such that the following three conditions are satisfied:*

$$\lambda\big(\mathbb{E}_{P_X^\star}[f(X)]-c\big)=0; \tag{12}$$

$$\forall x\in\Omega: i(x;P_X^\star,P_{Y|X})-\lambda\big(f(x)-\mathbb{E}_{P_X^\star}[f(X)]\big) \leq I(P_X^\star,P_{Y|X}); \tag{13}$$

$$i(x;P_X^\star,P_{Y|X})-\lambda\big(f(x)-\mathbb{E}_{P_X^\star}[f(X)]\big)=I(P_X^\star,P_{Y|X}), \tag{14}$$

where c *is a fixed constant as in* (3) *(i.e.,* $\mathbb{E}_{P_X}[f(X)]\leq c$*).*

Theorem 11. (Identity Theorem for Real-Analytic Functions [17]) *Let* $f,g:\mathbb{R}^n\to\mathbb{R}$ *be two real-analytic functions that agree on a set* $\mathcal{E}\subset\mathbb{R}^n$*. Then,* f *and* g *agree on* $\mathbb{R}^n$ *if one of the following conditions is satisfied:*

1) $\mathcal{E}$ *is an open set;*
2) $\mathcal{E}$ *is a set of a positive Lebesgue measure;*
3) *If* $n=1$*, then it suffices for* $\mathcal{E}$ *to be an arbitrary set with an accumulation point.*

We will also need the following definition.

Definition 7. A set $\mathcal{A}\subset\mathcal{X}$ is said to be *dense* in the set $\mathcal{X}$ if every element $x\in\mathcal{X}$ either belongs to $\mathcal{A}$ or is an accumulation 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 $\mathcal{X}$.

Theorem 12. (Properties of the Optimal Support) *Suppose that* $\Omega\subset\mathbb{R}^n$ *contains an open subset and let* $i(x;P_X^\star,P_{Y|X})$ *and* f *be non-constant, real analytic functions on* Ω*. Then,* $\mathcal{E}(P_X^\star)\subset\Omega\subset\mathbb{R}^n$ *is a nowhere dense set of* Ω *and is of Lebesgue measure zero. In addition, if* $n=1$*, then for every finite interval* $\mathcal{J}$*, the set* $\mathcal{E}(P_X^\star)\cap\mathcal{J}$ *is of finite cardinality.*

Proof: We first show that $\mathcal{E}(P_X^\star)$ is a nowhere dense set. Towards a contradiction, assume that $\mathcal{E}(P_X^\star)$ is not nowhere dense in Ω. Therefore, by Definition 7, there exists an open set $\mathcal{U}\subset\Omega$ such that $\mathcal{U}\cap\mathcal{E}(P_X^\star)$ is dense in Ω. By using (14), we have that

$$g(x)\coloneqq i(x;P_X^\star,P_{Y|X})-\lambda\big(f(x)-\mathbb{E}_{P_X^\star}[f(X)]\big)$$

is constant on $\mathcal{E}(P_X^\star)$ and therefore is constant on $\mathcal{U}\cap\mathcal{E}(P_X^\star)$. Since $\mathcal{U}\cap\mathcal{E}(P_X^\star)$ is dense in Ω by the properties of continuous functions (analytic functions are continuous), g is also constant on $\mathcal{U}$. Moreover, since $\mathcal{U}$ is an open set and g is analytic and constant on $\mathcal{U}$ by property 1) of Theorem 11, g must be constant on Ω. However, this leads to a contradiction as we assumed that g is non-constant on Ω. Therefore, $\mathcal{E}(P_X^\star)$ is a nowhere dense in Ω.

The conclusion that $\mathcal{E}(P_X^\star)$ has Lebesgue measure zero follows along similar lines by assuming that $\mathcal{E}(P_X^\star)$ is a set of positive measure and using property 2) of Theorem 11 to conclude that g must be constant on all of Ω. This again leads to a contradiction, which implies that $\mathcal{E}(P_X^\star)$ must have Lebesgue measure zero. ∎

Remark 3. Note that if f and $i(x;P_X^\star,P_{Y|X})$ are *orthogonally equivariant* (i.e., they only depend on $\|x\|$), then it is not difficult to see that $\mathcal{E}(P_X^\star)$ a union of concentric spheres. That is,

$$\mathcal{E}(P_X^\star)=\bigcup_j\mathcal{C}(r_j), \tag{15}$$

where $\mathcal{C}(r_j)\coloneqq\{x\in\mathbb{R}^n:\|x\|=r_j\}$, for some r_j. For example, this is the case if $P_{Y|X}=\mathcal{N}(x,I_n)$ with I_n the $n\times n$ identity matrix. The example in (15) shows that the cardinality of $\mathcal{E}(P_X^\star)$ is uncountable and that discrete inputs are in general not optimal. Theorem 12 can therefore generally not be improved in the sense that we cannot make statements about the cardinality of $\mathcal{E}(P_X^\star)$ if $\Omega=\mathbb{R}^n$ for $n>1$. Note, however, that the magnitude of $X\sim P_X^\star$ is discrete.

Remark 4. To the best of our knowledge, the approach taken in this section has never been used to obtain bounds on the number of mass points not even for a Gaussian channel with amplitude-constrained inputs (i.e,. $X\in[-A,A]$ for some $A>0$). An attempt to determine the position and the number of mass points was made in [18], where it was conjectured that the number of points increases by at most one. By using tools from complex analysis, one can show that if $x\mapsto i(x;P_X^\star,P_{Y|X})$ has a complex analytic extension to an open subset of $\mathbb{C}$ containing $[-A,A]$, then the number of mass points is given by

$$\big|\mathcal{E}(P_X^\star)\big|=\frac{1}{2\pi\mathrm{i}}\oint_\gamma\frac{i'(z;P_X^\star,P_{Y|X})}{i(z;P_X^\star,P_{Y|X})-I(P_X^\star,P_{Y|X})}\,\mathrm{d}z,$$

where γ is a regular closed curve that contains $[-A,A]$ and i' the derivative of i with respect to z. For details we refer to the extended version of this paper.

C. Some Remarks on Related Works

As mentioned at the beginning, the approach followed in this section was first presented in [16] in the context of scalar Gaussian channels with an amplitude or power-constrained input. In [19], the result was extended to the complex Gaussian case. The authors of [20] considered the Gaussian noise channel subject to Rayleigh fading, where channel state information is not available at the transmitter and the receiver and where channel input is subject to a power constraint. In particular, it was shown that the optimal input distribution is discrete albeit the number of mass points is countably infinite. In [21], similar result were obtained for power-constrained complex-valued channels with a rapidly varying phase.

In [22], the approach was applied to a large class of vector channels that are conditionally Gaussian (i.e., $P_{Y|X}$ is Gaussian) and where the input is constrained to an Euclidian ball and/or has finite power. There are also many works focusing on non-Gaussian additive noise channels. In [23], scalar additive noise channels with amplitude-constrained inputs are considered. The author provides sufficient conditions on the input

density to guarantee the optimal input distribution is discrete with finitely many points. For additive noise channels with arbitrary input constraints, the most general set of conditions under which the optimal input distribution is bounded or discrete can be found in [24]. For other results on additive channels with various input constraints, the interested reader is referred to [25]–[31]. Finally, it has to emphasized that the approach of this section has also been applied to non-additive noise channels [32], [33] as well as to multiuser channels [34].

V. Conclusion

In this work, we have focused on two optimization methods that follow ideas from the theories of linear and convex optimization. Of course there exist other approaches for finding capacity-achieving input distributions. For example, a promising approach is to connect, via I-MMSE type relationships [35]–[37], the theory of least-favorable prior distributions for estimation measures with the search for optimal input distributions. Such connections have been made in the context of Gaussian noise channels [38]–[40]. Another challenging future direction might be to evaluate how the approaches can be extended to multiuser settings such as the interference channel.

References

[1] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[2] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[3] T. Sutter, D. Sutter, P. M. Esfahani, and J. Lygeros, "Efficient approximation of channel capacities," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1649–1666, 2015.

[4] Y. Wu and S. Verdú, "The impact of constellation cardinality on Gaussian channel capacity," in *Proc. Allerton Conf. Commun., Control and Comp.*, Monticello, IL, 2010, pp. 620–628.

[5] 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 (ISIT)*, Aachen, Germany, 2017, pp. 1058–1062.

[6] M. Egan, S. M. Perlaza, and V. Kungurtsev, "Capacity sensitivity in continuous channels," INRIA Grenoble - Rhône-Alpes; Czech Technical University in Prague, Tech. Rep. RR-9012, 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01455640/document

[7] P. J. Huber, *Robust Statistics*. Wiley-Interscience, 1981.

[8] R. M. Dudley, *Real Analysis and Probability*. Cambridge University Press, 2002.

[9] D. G. Luenberger, *Optimization by Vector Space Methods*. John Wiley & Sons, 1997.

[10] G. Winkler, "Extreme points of moment sets," *Math. Oper. Res.*, vol. 13, no. 4, pp. 581–587, 1988.

[11] L. E. Dubins, "On extreme points of convex sets," *Math. Anal. Appl.*, vol. 5, no. 2, pp. 237–244, 1962.

[12] H. Witsenhausen, "Some aspects of convexity useful in information theory," *IEEE Trans. Inf. Theory*, vol. 26, no. 3, pp. 265–271, 1980.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

[14] Y. Wu, L. M. Davis, and R. Calderbank, "On the capacity of the discrete-time channel with uniform output quantization," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Seoul, Korea, 2009, pp. 2194–2198.

[15] T. Koch and A. Lapidoth, "At low SNR, asymmetric quantizers are better," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5421–5445, 2013.

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

[17] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*, 2nd ed. Springer, 2002.

[18] N. Sharma and S. Shamai (Shitz), "Transition points in the capacity-achieving distribution for the peak-power limited AWGN and free-space optical intensity channels," *Probl. Inf. Transm.*, vol. 46, no. 4, pp. 283–299, 2010.

[19] S. Shamai and I. Bar-David, "The capacity of average and peak-power-limited quadrature Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1060–1071, 1995.

[20] I. C. Abou-Faycal, M. D. Trott, and S. Shamai, "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1290–1301, 2001.

[21] M. Katz and S. Shamai, "On the capacity-achieving distribution of the discrete-time noncoherent and partially coherent AWGN channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2257–2270, 2004.

[22] T. H. Chan, S. Hranilovic, and F. R. Kschischang, "Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2073–2088, 2005.

[23] A. Tchamkerten, "On the discreteness of capacity-achieving distributions," *IEEE Trans. Inf. Theory*, vol. 50, no. 11, pp. 2773–2778, 2004.

[24] 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*, vol. 64, no. 2, pp. 1178–1198, 2018.

[25] A. ElMoslimany and T. M. Duman, "On the discreteness of capacity-achieving distributions for fading and signal-dependent noise channels with amplitude-limited inputs," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1163–1177, 2018.

[26] L. Zhang, H. Li, and D. Guo, "Capacity of Gaussian channels with duty cycle and power constraints," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1615–1629, 2014.

[27] H. Li, S. M. Moser, and D. Guo, "Capacity of the memoryless additive inverse Gaussian noise channel," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 12, pp. 2315–2329, 2014.

[28] A. Dytso, R. Bustin, H. V. Poor, and S. S. Shitz, "On additive channels with generalized Gaussian noise," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, 2017, pp. 426–430.

[29] A. Behboodi, G. Alirezaei, and R. Mathar, "On the discreteness of capacity-achieving distributions for the censored channel," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, 2017, pp. 1067–1071.

[30] L. R. Varshney, "Transporting information and energy simultaneously," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Toronto, Canada, 2008, pp. 1612–1616.

[31] O. Ozel and S. Ulukus, "AWGN channel under time-varying amplitude constraints with causal information at the transmitter," in *Proc. Asilomar Conf. Signals, Systems and Computers*, Pacific Grove, CA, 2011, pp. 373–377.

[32] S. Shamai, "Capacity of a pulse amplitude modulated direct detection photon channel," *IEE Proc. I (Commun., Speech and Vision)*, vol. 137, no. 6, pp. 424–430, 1990.

[33] M. V. Hegde, W. E. Stark, and D. Teneketzis, "On the capacity of channels with unknown interference," *IEEE Trans. Inf. Theory*, vol. 35, no. 4, pp. 770–783, 1989.

[34] B. Mamandipoor, K. Moshksar, and A. K. Khandani, "On the sum-capacity of Gaussian MAC with peak constraint," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Cambridge, MA, 2012, pp. 26–30.

[35] D. Guo, S. Shamai (Shitz), 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.

[36] S. Shamai (Shitz), "From constrained signaling to network interference alignment via an information-estimation perspective," *IEEE Inf. Theory Soc. Newsl.*, vol. 62, no. 7, pp. 6–24, 2012.

[37] J. Jiao, K. Venkat, and T. Weissman, "Relations between information and estimation in discrete-time Lévy channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3579–3594, 2017.

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

[39] A. Dytso, M. Goldenbaum, S. Shamai (Shitz), and H. V. Poor, "Upper and lower bounds on the capacity of amplitude-constrained MIMO channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, 2017, pp. 1–6.

[40] A. Dytso, R. Bustin, H. V. Poor, and S. Shamai (Shitz), "On the structure of the least favorable prior distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, 2018, submitted for publication.