Some Aspects of Totally Positive Kernels Useful in Information Theory

Semih Yagli†, Alex Dytso†, H. Vincent Poor†, Shlomo Shamai (Shitz)*,
† Princeton University, Princeton, NJ 08544, USA, Email: {syagli, adytso, poor}@princeton.edu
* Israel Institute of Technology, Technion City, Haifa 32000, Israel, Email: sshlomo@ee.technion.ac.il

***Abstract*—This paper introduces totally positive kernels and Pólya type distributions to information theory. In particular, it is shown that the variational diminishing property of Pólya type distributions, which is captured by the Oscillation Theorem, can be used to characterize the structure of capacity-achieving distributions for a large class of channels.**

I. Introduction

In his seminal paper [1], Shannon has characterized the fundamental limit of communication over a noisy channel by the following optimization problem:

$$C = \max_{X \in \Omega_X} I(X;Y), \tag{1}$$

where $I(X;Y)$ is the mutual information between the channel input X, and the channel output Y, and Ω_X is a constraint on the input, e.g., $\Omega_X = [-A, A]$.

Even in the simplest settings, finding the optimizer in (1) is often too unwieldy, and in fact this feat can only be accomplished in certain special cases of Ω_X. Under the framework that Ω_X is a compact interval, if the channel transition kernel $P_{Y|X}$ satisfies a mild set of regularity conditions, this work proposes a general method that is able to characterize the structure of the optimal input distribution by showing that the optimal input distribution is discrete with finitely many mass points. The method is shown to work for a wide family of channel transition probabilities $P_{Y|X}$ termed Pólya type distributions.

A. Contributions and Paper Outline

The main actors in this paper, the totally positive kernels and the Pólya type distributions, are described in Section II. While it also contains several important theorems regarding various properties of these type of kernels, the most important theorem in this paper, namely the Oscillation Theorem, along with the definitions of the two novel concepts, a transform based on totally positive kernels and its inverse transform,[1] are presented in Section II, as well.

The role of the Pólya type distributions and the Oscillation Theorem in information theory are described in Section III. In particular, under very mild conditions on the random transformation $P_{Y|X}$, it is shown that if the constraint set Ω_X is a compact interval, then the optimizing input X in (1) is a discrete random variable whose number of mass points cannot exceed the number of zeros of a downward shifted output probability density function (pdf). Last but not least, Section III concludes with the discussion on how such results can be extended to channels with more abstract output spaces.

[1]Presuming the inverse transform exists. Existence conditions for the inverse of the transform that is based on Pólya type functions are part of our ongoing research.

B. Past Work

In this subsection, we give a brief overview of the available methods for finding capacity-achieving distributions. An in-detail summary of this topic can be found in [2].

In the case when channel input and output spaces have finite cardinalities, capacity-achieving distributions can be obtained numerically by using the Blahut-Arimoto algorithm [3], [4]. Specifically, in the case when the channel transition probability is symmetric (i.e., all rows of the channel transition matrix are permutations of each other) it is well-known that the equiprobable distribution over the input alphabet achieves the channel capacity [5, Chapter 7.2]. Indeed, for channels with finite alphabets, Gallager in [6, Corollary 3, p. 97] has shown that the smallest number of channel input symbols that can be used with nonzero probability to achieve capacity must be less than the cardinality of the output space.

For channels whose input or output (or both) alphabet has infinite cardinality (either countable or uncountable) the problem of finding a capacity-achieving distribution takes a form of an infinite dimensional optimization and is often too difficult to solve. However, several results are known regarding the structure of the support of capacity-achieving distributions. For instance, in the case when the output space is finite and the input space is an arbitrary compact set, Witsenhausen [7], using Dubins' hyperplane theorem, has generalized the result of Gallager by showing that a capacity-achieving distribution is discrete with the number of mass points upper bounded by the cardinality of the output space. This bound is often tight. Regretfully, Witsenhausen's elegant approach does not work for the arbitrary output alphabet case. Luckily, the approach presented here is able to extend Witsenhausen's result to more generic output alphabet cases.

The most general approach for finding the structure of capacity-achieving distributions, which works for arbitrary alphabets, employs convex optimization methods [6, Theorem 4.5.1]; see also [2]. The caveat, however, is that such an approach usually needs additional tools from complex analysis and the theory of analytic functions. This marriage of convex and complex analyses has been used successfully to show

978-1-7281-0922-0/19/$31.00 ©2019 IEEE

that the capacity-achieving distributions are discrete for several practically relevant channels such as the Gaussian noise channel with an input amplitude constraint [8], Poisson channel with an input peak-intensity constraint [9], Rayleigh fading channel with an input power constraint [10], and arbitrary additive channels where the noise pdf is an analytic function [11]. However, a drawback of this technique is that the proofs are inherently non-constructive, and the obtained results only show discreteness without providing any type of bounds on the support size of capacity-achieving distributions. One of the aims of this paper is to fill this theoretical gap by providing a constructive method based on the properties of Pólya type distributions. Not only are the introduced tools strong enough to show that a capacity-achieving distribution is discrete with finite support, but they are also able to provide concrete upper bounds on the number of elements in that support.

Finally, there are several ad-hoc methods for finding capacity-achieving distributions. One such method is the maximum entropy principle introduced by Shannon in [1] who showed the optimality of a Gaussian input distribution over the additive Gaussian noise channel with a power constraint on the input. The maximum entropy principle has also been successfully applied to find the capacity-achieving distribution of an additive exponential noise channel with a first moment constraint on the input [12]. Other ad-hoc methods use the notion of stochastic dominance or estimation theoretic techniques. For these and other methods the interested readers are referred to [2].

C. Notation

Throughout the paper, we denote the distribution of a random variable X by P_X. Moreover, we say that a point x is in the support, denoted by $\mathsf{supp}(P_X)$,[2] of the distribution P_X if for every open set $\mathcal{O} \ni x$ we have that $P_X(\mathcal{O}) > 0$.

The number of zeros of a function $f\colon \mathbb{R} \to \mathbb{R}$ on the interval $\mathbb{I}$ is denoted by $\mathrm{N}(\mathbb{I}, f)$. Similarly, if $f\colon \mathbb{C} \to \mathbb{C}$ is a function on the complex domain, $\mathrm{N}(\mathcal{D}, f)$ denotes the number of its zeros within the region $\mathcal{D}$. The interior of the set $\mathcal{O}$ is denoted by $\mathring{\mathcal{O}}$.

Finally, while the mutual information between X and Y is denoted by $I(X;Y)$, the entropy of a discrete random variable X is denoted by $H(X)$ and the differential entropy of a continuous random variable X is denoted by $h(X)$.

II. Main Tool: Totally Positive Kernels and Pólya Type Distributions

The main tool used in this paper is the variational diminishing property of Pólya type distributions introduced by Karlin [13] in the context of the statistical decision theory. Before introducing this key property, we shall describe some important concepts, namely totally positive kernels and Pólya type distributions, required in our analysis.

[2]Also known as "points of increase of P_X" or "spectrum of P_X."

A. Totally Positive Kernels

Definition 1 (Totally Positive Kernel[3])**.** *Given arbitrary real subsets, $\mathcal{S}_1$ and $\mathcal{S}_2$, a function $f\colon \mathcal{S}_1 \times \mathcal{S}_2 \to \mathbb{R}$ is said to be a* totally positive kernel of order n *if*

$$\det \begin{bmatrix} f(x_1,y_1) & \cdots & f(x_1,y_m) \\ \vdots & \ddots & \vdots \\ f(x_m,y_1) & \cdots & f(x_m,y_m) \end{bmatrix} \geq 0 \tag{2}$$

for all $1 \leq m \leq n$, and for all $x_1 < \cdots < x_m \in \mathcal{S}_1$ and $y_1 < \cdots < y_m \in \mathcal{S}_2$. A function f is called a totally positive kernel *if it is a totally positive kernel of order n for every choice of $n \in \mathbb{N}$. Moreover, if* (2) *holds with strict inequality, then the function f is called a* strictly totally positive kernel of order n *and a* strictly totally positive kernel*, respectively.*

In what follows, the class of *strictly* totally positive kernels of order n and the class of strictly totally positive kernels are denoted by $\mathcal{T}_n$ and $\mathcal{T}_\infty$, respectively.

From Definition 1, $f(\cdot,\cdot) \in \mathcal{T}_1$ if and only if

$$f(x,y) > 0, \tag{3}$$

for all (x,y). Moreover, $f(\cdot,\cdot) \in \mathcal{T}_2$ if and only if the ratio $f(x_1,y)/f(x_2,y)$ is strictly monotone decreasing in y for $x_1 < x_2$.

Further imposing symmetry and positive (semi-)definiteness in the definition of totally positive kernels, one recovers the definition of positive (semi-)definite kernels.

Definition 2 (Positive (Semi-)Definite Kernel [15])**.** *A symmetric function $f : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is said to be a positive definite kernel if the matrix*

$$\begin{bmatrix} f(x_1,x_1) & \cdots & f(x_1,x_m) \\ \vdots & \ddots & \vdots \\ f(x_m,x_1) & \cdots & f(x_m,x_m) \end{bmatrix}$$

is positive-definite for all $m \in \mathbb{N}$ and for all $x_1 \neq \cdots \neq x_m \in \mathcal{S}$.

Employing the relationship between positive definite matrices and their determinants, as stated below, it is easy to see that the class of strictly totally positive kernels contains that of positive definite kernels.

Lemma 1. *Let $\mathcal{K}$ and $\mathcal{T}_\infty$ denote the set of all positive definite kernels and the set of all strictly totally positive kernels, respectively. Then, $\mathcal{K} \subset \mathcal{T}_\infty$.*

Positive definite kernels are often encountered in machine learning applications [15]. However, while the symmetry assumption made in Definition 2 is common for many machine learning applications, it is too restrictive for the purposes of this paper.

[3]Definition 1 is a generalization of functions introduced by Pólya in [14] where only the shift families, i.e., $f(x,y) = g(x-y)$, were considered.

The following theorem [16, Lemma 5] is a useful tool in checking whether a given function is a member of $\mathcal{T}_n$.

Theorem 1. *Let μ be a σ-finite measure. For $p(\cdot,\cdot) \in \mathcal{T}_n$ and $q(\cdot,\cdot) \in \mathcal{T}_m$, suppose*

$$r(x,y) = \int p(x,t)q(t,y)\mathrm{d}\mu(t). \tag{4}$$

Then, $r(\cdot,\cdot) \in \mathcal{T}_{\min(n,m)}$.

Another useful instrument in verifying whether a function f is a member of $\mathcal{T}_\infty$ is the following theorem [17].

Theorem 2. *$\mathcal{T}_\infty$ satisfies the following:*

1) (Closure Under Positive Linear Combinations) *Given $\{f_1,\ldots,f_n\} \subset \mathcal{T}_\infty$ and $\alpha_i \geq 0, i = 1,\ldots,n$, it follows that $\sum_{i=1}^n \alpha_i f_i \in \mathcal{T}_\infty$.*
2) (Closure Under Limits) *For any (x,y), suppose that $f_\infty(x,y) = \lim_{i\to\infty} f_i(x,y)$ where $f_i \in \mathcal{T}_\infty$. Then, $f_\infty \in \mathcal{T}_\infty$.*

B. Pólya Type Distributions

Building upon the familiarity with the totally positive kernels established in the previous section, we define Pólya type distributions as follows:

Definition 3 (Strictly Pólya Type Distribution)**.** *Let $\mathcal{S}$ be an arbitrary real subset and $\mathbb{I}$ be an open interval. A distribution $P(\cdot|x)$ on the set $\mathcal{S}$ that depends on the parameter $x \in \mathbb{I}$ is said to be a* strictly Pólya type-n distribution *if*

$$P(y|x) = \beta(x)\int_{-\infty}^{y} f(t,x)\mathrm{d}\nu(t), \tag{5}$$

where ν is a σ-finite measure, $\beta(\cdot) > 0$ is the normalization constant and $f(\cdot,\cdot) \in \mathcal{T}_n$. If $P(\cdot|x)$ is a strictly Pólya type-n distribution for all $n \in \mathbb{N}$, then $P(\cdot|x)$ is said to be a strictly Pólya type-∞ distribution.

To make it easier to refer to them, in what follows the class of strictly Pólya type-n and the class of strictly Pólya type-∞ functions are denoted by $\mathcal{P}_n$ and $\mathcal{P}_\infty$, respectively.

Immediate from the above definition, observe that a continuous distribution is a member of $\mathcal{P}_\infty$ if and only if its pdf is an element of $\mathcal{T}_\infty$. Similarly, a discrete distribution is in $\mathcal{P}_\infty$ if and only if its probability mass function (pmf) is contained in $\mathcal{T}_\infty$.

Note that many distributions are members of $\mathcal{P}_\infty$. As an example, consider the large family of single-parameter exponential distributions whose cumulative distribution functions (cdfs) can be written as

$$F(y|x) = \lambda(x)\int_{-\infty}^{y} \mathrm{e}^{xt}\mathrm{d}\mu(t), \tag{6}$$

where $\lambda(x) > 0$, $f(x,t) = \mathrm{e}^{xt} \in \mathcal{T}_\infty$, and μ is some σ-finite measure. While a non-exhaustive list of Pólya type-∞ distributions is given in Table I, some other examples from $\mathcal{P}_\infty$ can be found in [13], [16] and [18].

TABLE I: Common Distributions that are Pólya type-∞. We denote the pdf's by $f(\cdot|x)$; and the pmf's by $p(\cdot|x)$.

Distribution Name	PDF or PMF
Gaussian	$f(y\|x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(y-x)^2}{2}},\ y \in \mathbb{R},\ x \in \mathbb{R}$
Laplace	$f(y\|x) = \frac{1}{2}\mathrm{e}^{-\|y-x\|},\ y \in \mathbb{R},\ x \in \mathbb{R}$
Poisson	$p(y\|x) = \frac{x^y\mathrm{e}^{-x}}{y!},\ y \in \mathbb{N}\cup\{0\},\ x \geq 0$
Noncentral Chi-Square	$f(y\|x) = \frac{1}{2}\mathrm{e}^{-\frac{y+x}{2}}\left(\frac{y}{x}\right)^{\frac{k-2}{4}} \mathsf{I}_{\frac{k}{2}-1}(\sqrt{xy})$, $y \geq 0, x \geq 0, k > 0$
Binomial	$p(y\|x) = \binom{n}{y}x^y(1-x)^{n-y}$, $y \in \{0,1,\ldots,n\}, x \in [0,1]$
Gamma	$f(y\|x) = \frac{1}{x^k}y^{k-1}\mathrm{e}^{-\frac{y}{x}},\ y > 0,\ x > 0,\ k > 0$

Still, not every probability distribution is a member of $\mathcal{P}_\infty$. A non-example is the Cauchy distribution, whose density is given by

$$f(y|x) = \frac{1}{\pi}\frac{1}{1+(y-x)^2}. \tag{7}$$

Clearly, $f \in \mathcal{T}_1$. To see that $f \notin \mathcal{T}_n$ for $n \geq 2$, observe that choosing $(x_1,x_2,y_1,y_2) = (1,2,3,4)$ results in

$$\det\begin{bmatrix} f(y_1|x_1) & f(y_2|x_1) \\ f(y_1|x_2) & f(y_2|x_2)\end{bmatrix} = -\frac{1}{(10\pi)^2} < 0, \tag{8}$$

violating the condition in (2).

C. Oscillation Theorem

An important feature of the Pólya type distributions is their variational diminishing property, which is captured by the Oscillation Theorem. The following definition sets the stage for this property.

Definition 4 (Number of Sign Changes of a Function)**.** *The number of sign changes of a function ξ on the set $\mathcal{S}$ is given by*

$$\mathscr{S}(\mathcal{S},\xi) = \sup_{m\in\mathbb{N}}\left\{\sup_{\substack{y_i\in\mathcal{S}:\\ y_1<\cdots<y_m}} \mathscr{N}\{\xi(y_i)\}_{i=1}^m\right\}, \tag{9}$$

where $\mathscr{N}\{\xi(y_i)\}_{i=1}^m$ is the number of sign changes of the sequence $\{\xi(y_i)\}_{i=1}^m$.

Proven in [16], the following theorem is the main tool in connecting the number of zeros of the output pdf $f_{Y^\star}$ (or pmf $p_{Y^\star}$ in the case of discrete output alphabet) to the number of mass points of a capacity-achieving input distribution $P_{X^\star}$.

Theorem 3 (Oscillation Theorem)**.** *Given an arbitrary set $\mathcal{S} \subset \mathbb{R}$ and an open interval $\mathbb{I}$, let $p(\cdot|\cdot)\colon \mathcal{S}\times\mathbb{I} \to \mathbb{R}$ be a strictly Pólya type-∞ distribution. For an arbitrary but fixed y, suppose $p(y|\cdot)\colon \mathbb{I} \to \mathbb{R}$ is an n-times differentiable function.*

Assume that μ is a measure on $\mathcal{S}$, and let $\xi\colon \mathcal{S} \to \mathbb{R}$ be a function with $\mathscr{S}(\mathcal{S},\xi) = n$. For $x \in \mathbb{I}$, define

$$\Xi(x) = \int \xi(y)p(y|x)\mathrm{d}\mu(y). \tag{10}$$

If $\Xi\colon \mathbb{I} \to \mathbb{R}$ is an n-times differentiable function, then either of following statements holds:

1) *$\Xi \not\equiv 0$, and $\mathscr{S}(\mathcal{S},\Xi) \leq \mathrm{N}(\mathbb{I},\Xi) \leq \mathscr{S}(\mathcal{S},\xi) = n$; or*
2) *$\Xi \equiv 0$. This holds if and only if the $\mathsf{supp}(\mu)$ is a subset of zeros of $\xi(\cdot)$.*

D. A Transform Based on Totally Positive Kernels

This section is devoted to the definition of a transform based on totally positive kernels.

Definition 5. *Let $f\colon \mathbb{I} \to \mathbb{R}$ and let $p(\cdot|\cdot) \in \mathcal{T}_n$. Then, a* Pólya transform *of a function f with respect to kernel $p(\cdot|\cdot)$ is defined as*

$$\mathscr{P}[f](x) = \int f(y)p(y|x)\mathrm{d}\mu(y). \tag{11}$$

An inverse of this transform, provided that it exists, is denoted by $\mathscr{P}^{-1}$.

While the definition of this transform is clear-cut, the existence of its inverse is part of our ongoing research. Although, it should be noted that there are cases where the inverse of the above transform is a non-issue. For example, if $p(\cdot|x)$ is Gaussian pdf with mean x, then the above transform becomes what is known as the Weierstrass transform, which has a well-defined inverse transform [19].

III. Application in Information Theory: Structure of Capacity-Achieving Distributions

In this section we show how the tools presented in Section II can be used to find properties of capacity-achieving input distributions for a large class of channels.

Consider a memoryless point-to-point channel with the input space $\Omega_X \subset \mathbb{R}$, the output space $\Omega_Y \subset \mathbb{R}$ and the channel transition probability $P_{Y|X}$. For ease of presentation, we assume that probability measure $P_{Y|X=x}$ has a probability density function denoted by $f_{Y|X}(\cdot|x)$. We note, however, that the results of this section generalize to more abstract cases.

Suppose that the set of all feasible inputs, $\Omega_X \subset \mathbb{R}$, is a compact interval. Observe that the compact interval assumption on Ω_X is not only ubiquitous, but it also carries practical relevance. For example, if the channel is Gaussian, i.e., $P_{Y|X=x} \sim \mathcal{N}(x,1)$, and $\Omega_X = [-A,A]$ we have the so-called peak-power constrained additive Gaussian channel [8], and if the channel is Poisson, i.e., $P_{Y|X=\lambda} \sim \mathrm{Poi}(\lambda)$, and $\Omega_X = [0,A]$ we have the so-called intensity constrained Poisson channel [9].

It is well known that the capacity expression for channels with such constraints is given by

$$C = \max_{X\in\Omega_X} I(X;Y), \tag{12}$$

where it is assumed that a capacity-achieving $X^\star \in \Omega_X$ exists. The situations in which a capacity-achieving distribution may not exist, in which case the maximum in (12) should be replaced by the supremum, are outside the scope of our present treatment.

Using elementary tools from convex optimization, the maximizing input $X^\star$ in (12) must satisfy the following necessary and sufficient conditions [2].

Lemma 2. *$X^\star \sim P_{X^\star}$ is a capacity-achieving distribution in* (12) *if and only if the following two conditions are satisfied:*

$$\forall x \in \Omega_X\colon i(x;P_{X^\star},P_{Y|X}) \leq \max_{X\in\Omega_X} I(X;Y), \tag{13}$$

$$\forall x \in \mathsf{supp}(P_{X^\star})\colon i(x;P_{X^\star},P_{Y|X}) = \max_{X\in\Omega_X} I(X;Y), \tag{14}$$

where

$$i(x;P_X,P_{Y|X}) = \mathbb{E}\left[\log\left(\frac{f_{Y|X}(Y|X)}{f_Y(Y)}\right)\middle| X = x\right]. \tag{15}$$

For our analysis, the following definition is also required:

$$h(Y|X=x) = \int \log\left(\frac{1}{f_{Y|X}(y|x)}\right) f_{Y|X}(y|x)\mathrm{d}y. \tag{16}$$

Under very mild conditions, the next result shows that for a Pólya type-∞ channel with a compact interval constraint Ω_X on the input, a capacity-achieving distribution is discrete with a number of mass points not exceeding the number of zeros of a downward shifted output pdf.

Theorem 4. *Suppose that the following conditions hold:*

1) *$P_{Y|X}$ is Pólya type-∞;*
2) *$h(Y|X=\cdot)$ has an inverse Pólya transform (see Definition 5); and*
3) *There exists an $x \in \Omega_X$ such that*

$$i(x;P_{X^\star},P_{Y|X}) < \max_{X\in\Omega_X} I(X;Y). \tag{17}$$

Then,

$$\begin{aligned} &|\mathsf{supp}(P_{X^\star})| \\ &\quad\leq \mathrm{N}\left(\Omega_Y, f_{Y^\star}(\cdot) - \mathrm{e}^{-\mathscr{P}^{-1}[h(Y|X=\cdot)](\cdot)-C}\right) + 2 \qquad (18)\\ &\quad< \infty, \qquad (19)\end{aligned}$$

provided that $f_{Y^\star}(\cdot) - \mathrm{e}^{-\mathscr{P}^{-1}[h(Y|X=\cdot)](\cdot)-C}$ has finitely many zeros on Ω_Y.

Proof. Let

$$\Xi(x;P_{X^\star},P_{Y|X}) = i(x;P_{X^\star},P_{Y|X}) - C. \tag{20}$$

Suppose that $\mathrm{N}\left(\Omega_X,\Xi(\cdot;P_{X^\star},P_{Y|X})\right) < \infty$. An immediate consequence of Lemma 2 is the fact that if $x \in \mathsf{supp}(P_{X^\star})$ then $i(x;P_X,P_{Y|X}) - C = 0$. In other words,

$$|\mathsf{supp}(P_{X^\star})| \leq \mathrm{N}\left(\Omega_X,\Xi(\cdot;P_{X^\star},P_{Y|X})\right). \tag{21}$$

Next, observe that $\Xi(\cdot; P_{X^\star}, P_{Y|X})$ can be re-written as follows:

$$\begin{aligned}
&\Xi(x; P_{X^\star}, P_{Y|X}) \\
&\quad = \int \log \frac{1}{f_Y(y)} f_{Y|X}(y|x)\mathrm{d}y - C - h(Y|X=x) \quad (22)\\
&\quad = \int \log \frac{\mathrm{e}^{-\mathscr{P}^{-1}[h(Y|X=\cdot)](y)-C}}{f_Y(y)} f_{Y|X}(y|x)\mathrm{d}y \quad (23)\\
&\quad = \int \xi(y) f_{Y|X}(y|x)\mathrm{d}\mu(y), \quad (24)
\end{aligned}$$

where in (23) we have used Assumption 2 that the inverse Pólya transform of function $h(Y|X=\cdot)$ exits; and in (24)

$$\xi(y) = \log \frac{1}{f_{Y^\star}(y)} - \mathscr{P}^{-1}\left[h(Y|X=\cdot)\right](y) - C. \quad (25)$$

Now, using the fact that the $f_{Y|X}$ is a Pólya Type-∞, and resuming from (21)

$$\begin{aligned}
&|\mathsf{supp}(P_{X^\star})| \\
&\quad \le \mathrm{N}\left(\Omega_X, \Xi(x; P_{X^\star}, P_{Y|X})\right) \quad (26)\\
&\quad \le \mathrm{N}\left(\mathring{\Omega}_X, \Xi(x; P_{X^\star}, P_{Y|X})\right) + 2 \quad (27)\\
&\quad \le \mathscr{S}(\Omega_Y, \xi) + 2 \quad (28)\\
&\quad \le \mathrm{N}(\Omega_Y, \xi) + 2 \quad (29)\\
&\quad = \mathrm{N}\left(\Omega_Y, f_{Y^\star}(\cdot) - \mathrm{e}^{-\mathscr{P}^{-1}[h(Y|X=\cdot)](\cdot)-C}\right) + 2, \quad (30)
\end{aligned}$$

where (27) follows from restricting attention to the interior of Ω_X and accounting for possible zeros at the two boundary points; (28) follows from Theorem 3 where we have used Assumption 3 to eliminate the possibility of $\Xi(\cdot; P_{X^\star}, P_{Y|X}) \equiv 0$; (29) follows because the number of zeros is an upper bound on the number of sign changes; and (30) follows by observing that $\xi(y) = 0$ if and only if $f_{Y^\star}(y) - \mathrm{e}^{-\mathscr{P}^{-1}[h(Y|X=\cdot)](y)-C} = 0$.

Lastly, suppose $\mathrm{N}\left(\Omega_X, \Xi(\cdot; P_{X^\star}, P_{Y|X})\right) = \infty$. Then, (24) and Theorem 3 enforce that $f_{Y^\star}(\cdot) - \mathrm{e}^{-\mathscr{P}^{-1}[h(Y|X=\cdot)](\cdot)-C}$ cannot have finitely many zeros, causing a contradiction. ■

Several comments are in order.

I) An inverse Pólya transform always exists for an additive Pólya type-∞ channel with finite noise differential entropy. That is, suppose that the channel input-output relation is given by $Y = X + Z$, where the noise random variable Z satisfies $|h(Z)| < \infty$, and $\mathsf{supp}(Z) = \mathbb{R}$. In this case, $h(Y|X=\cdot) = h(Z)$ is a constant, and $\mathscr{P}^{-1}\left[h(Y|X=\cdot)\right](\cdot) \equiv h(Z)$. As a result, (18) becomes

$$|\mathsf{supp}(P_{X^\star})| \le \mathrm{N}\left(\Omega_Y, f_{Y^\star} - \mathrm{e}^{-h(Z)-C}\right) + 2. \quad (31)$$

II) Suppose, in addition to the conditions in Item I), that the output pdf is differentiable. Then, by Rolle's Theorem from elementary calculus, we can loosen the upper bound in (18) as

$$|\mathsf{supp}(P_{X^\star})| \le \mathrm{N}(\Omega_Y, f'_{Y^\star}) + 3. \quad (32)$$

Although looser, the benefit of (32) over (18) is that it reduces the dependence on the channel capacity C, which is typically unknown, because, in general, the maximizing distribution is unknown.

III) The assumption that $f_{Y^\star}(\cdot) - \mathrm{e}^{-\mathscr{P}^{-1}[h(Y|X=\cdot)](\cdot)-C}$ has finitely many zeros on Ω_Y is not a restrictive assumption. For example, it is satisfied when the channel is additive and the output pdf is an analytic function[4] on $\mathbb{R}$. Indeed, this is the case for an additive Gaussian channel with peak power constraint $\Omega_X = [-A, A]$. In this case, using the properties of the zeros of analytic functions, we arrive at

$$|\mathsf{supp}(P_{X^\star})| < \infty. \quad (33)$$

In particular, the bound in (33) gives an alternative method of proving Smith's result in [8] where it was shown that for a Gaussian noise channel the maximizing input distribution is discrete with finitely many points. In fact, due to limitations of the technique Smith used in his proof, unknown before was a firm upper bound on $|\mathsf{supp}(P_{X^\star})|$, which is now shown in our recent work to be $\mathrm{O}(A^2)$; see [20].

ACKNOWLEDGEMENT

This work was supported in part by the U. S. National Science Foundation under Grants CCF-093970 and CCF-1513915; and by the European Union's Horizon 2020 Research and Innovation Programme, grant agreement no. 694630.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 379-423, 623-656, 1948.

[2] A. Dytso, M. Goldenbaum, H. V. Poor, and S. Shamai (Shitz), "When are discrete channel inputs optimal? - Optimization techniques and some new results," in *Proc. Conf. on Inf. Sci. and Sys.*, Princeton, NJ, USA, March 2018, pp. 1–6.

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

[4] 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.

[5] T. Cover and J. Thomas, *Elements of Information Theory: Second Edition*. Wiley, 2006.

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

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

[8] J. G. Smith, "On the Information Capacity of Peak and Average Power Constrained Gaussian Channels," PhD dissertation, University of California, 1969.

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

[10] 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.

[11] 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.

[12] S. Verdú, "The exponential distribution in information theory," *Problemy Peredachi Informatsii*, vol. 32, no. 1, pp. 100–111, 1996.

[4] A function f that has a power series representation on an open set D is said to be analytic on D.

[13] S. Karlin, "Decision theory for Pólya type distributions. Case of two actions, I," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.
[14] G. Pólya, "Über annäherung durch polynome mit lauter reellen wurzeln," *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, vol. 36, no. 1, pp. 279–295, 1913.
[15] G. E. Fasshauer, "Positive definite kernels: Past, present and future." [Online]. Available: http://www.math.iit.edu/ fass/PDKernels.pdf
[16] S. Karlin, "Pólya type distributions, II," *The Ann. Math. Stat.*, vol. 28, no. 2, pp. 281–308, 1957.
[17] ——, *Total Positivity*. Stanford University Press, 1968, vol. 1.
[18] S. Karlin and H. Rubin, "The theory of decision procedures for distributions with monotone likelihood ratio," *The Annals of Mathematical Statistics*, pp. 272–299, 1956.
[19] A. I. Zayed, *Handbook of Function and Generalized Function Transforms*. CRC press, 1996.
[20] A. Dytso, S. Yagli, H. V. Poor, and S. Shamai, "Capacity achieving distribution for the amplitude constrained additive Gaussian channel: An upper bound on the number of mass points," 2019. [Online]. Available: http://www.princeton.edu/∼adytso/papers/ISIT2019.pdf