Capacity-Achieving Probability Measure for Conditionally Gaussian Channels With Bounded Inputs

Terence H. Chan, *Member, IEEE*, Steve Hranilovic, *Member, IEEE*, and Frank R. Kschischang, *Senior Member, IEEE*

***Abstract*—A conditionally Gaussian channel is a vector channel in which the channel output, given the channel input, has a Gaussian distribution with (well-behaved) input-dependent mean and covariance. We study the capacity-achieving probability measure for conditionally Gaussian channels subject to bounded-input constraints and average cost constraints. Many practical communication systems, including additive Gaussian noise channels, certain optical channels, fading channels, and interference channels fall within this framework. Subject to bounded-input constraint (and average cost constraints), we show that the channel capacity is achievable and we derive a necessary and sufficient condition for a probability measure to be capacity achieving. Under certain conditions, the capacity-achieving measure is proved to be discrete.**

***Index Terms*—Bounded-input constraint, capacity-achieving measure, conditionally Gaussian channel, optical channel, Rayleigh-fading channel.**

I. Introduction

WE STUDY the capacity-achieving probability measure under boundedness constraint for conditionally Gaussian channels, a class of vector channels whose conditional output distribution, given the channel input, is Gaussian with input-dependent mean and, in general, input-dependent covariance. Classical additive white Gaussian noise channels, multiple-input multiple-output (MIMO) Rayleigh-fading channels, certain interference channels, and certain optical channels with signal-dependent noise are examples of communication channel models that fall within this framework.

Of course, determining the capacity of a channel subject to various input constraints is a classical problem of information theory. Shannon demonstrated that in the case of a scalar additive Gaussian noise channel subject to an average power constraint, the capacity-achieving distribution is Gaussian. In practice, however, this source distribution is not realizable due to its unbounded amplitude. A limit on the peak amplitude, as well as on the average power, is necessary to more accurately reflect the physical limitations present in most practical communication systems.

In [1], Smith studied the capacity of a scalar Gaussian channel subject both to an average power constraint and to a peak power constraint. The capacity-achieving distribution was shown to be discrete, with a finite number of probability mass points. Using a similar approach, Shamai and his colleagues [2]–[5] showed that the capacity-achieving measures are also discrete for many other channels, including Poisson channels, quadrature Gaussian channels, Rayleigh-fading channels and so on. In particular, since the publication of [4], there has been a wide interest in this area, and many channels have been found to have discrete capacity-achieving measures. In Table I, we list some channel models that we are aware of having been shown to have discrete capacity-achieving measures.

In this paper, we extend the work of Smith and Shamai to study the capacity-achieving measure for conditionally Gaussian channels. We have organized the paper as follows. In Section II, we present the channel model and we specify the channel input constraints. In Section III, we provide the main results of this paper. Subject to a bounded-input constraint and average cost constraints, we show that the channel capacity is achievable, and derive a necessary and sufficient condition for a probability measure to be capacity achieving. Using this necessary and sufficient condition, we propose an algorithm to find the capacity-achieving measure of a signal-dependent optical channel, which is traditionally difficult to analyze. Using an approach similar to that of [1], we prove that, under suitable criteria, the capacity-achieving measure is discrete. In the special case when the conditionally Gaussian channel is constant, we further prove that the capacity-achieving measures approach a Gaussian distribution if the bounded-input constraint is relaxed. The proofs for our main results rely on a collection of intermediate propositions and lemmas, which are stated and proved in Appendix B. In Section IV, we apply our results to analyze several practical channels, including certain optical channels, fading channels, interference channels, and parallel Gaussian channels. Finally, we provide some conclusions in Section V.

II. System Model

A. Notation

In this paper, we adopt the following notation. For positive integers N, M, and K, let $\mathsf{N} = \{1, \ldots, N\}$, $\mathsf{M} = \{1, \ldots, M\}$,

Manuscript received September 17, 2003; revised December 7, 2004. The work of T. H. Chan was supported in part by a Croucher Foundation Fellowship. The material in this paper was presented at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004.

T. H. Chan is with the Department of Computer Science, University of Regina, Regina, SK S4S 0A2, Canada (e-mail: terence@cs.uregina.ca).

S. Hranilovic is with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON L8S 4K1, Canada (e-mail: hranilovic@mcmaster.ca).

F. R. Kschischang is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: frank@comm.utoronto.ca).

Communicated by A. Lapidoth, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2005.847707

TABLE I
A LIST OF CHANNELS WITH DISCRETE CAPACITY-ACHIEVING MEASURES

Channel model	Constraints	References
Scalar additive Gaussian	**PP**, **AP**[1]	Smith [1] 1971
Additive noise with piecewise constant noise density	**PP**	Oettli [6] 1974
Poisson	**PP**, **AP**	Shamai [2] 1990
Quadrature Gaussian	**PP**, **AP**	Shamai *et al.* [3] 1995
Scalar additive heavy-tailed noise	**AP**	Das [7] 2000
Rayleigh fading	**AP**	Abou-Faycal *et al.* [4] 2001
Non-coherent scalar AWGN	**AP**	Katz *et al.* [5] 2002
Additive vector Gaussian	**PP**, **AP**	Palanki [8] 2002
Single antenna Rayleigh block fading	**AP**	Palanki [8] 2002
General scalar fading	higher-moment constraint[2]	Palanki [8] 2002
Non-coherent Rician fading	**AP**	Gursoy *et al.* [9] 2003
Non-coherent block independent AWGN	**AP**	Nuriyev *et al.* [10] 2003
Partially coherent AWGN	**AP**	Hou *et al.* [11] 2003
UIUO[3]	**PP**	Huang *et al.* [12] 2003
General scalar additive	**PP**	Tchamkerten [13] 2004

[1] **PP** and **AP** stand for "peak power" and "average power" respectively.

[2] A higher moment constraint is an average cost constraint whose cost function is in the form: $x^\alpha - \gamma$ where $\alpha > 2$.

[3] This is a name we gave to the channel described in [12]. Here, UIUO stands for "Unbounded-Input Unbounded-Output." A UIUO channel is a scalar channel such that for any $c > 0, \lim_{|x|\to\infty} \Pr(|Y| < c|X = x) = 0$ where X and Y are channel input and output, respectively. In other words, if the channel input goes to infinity, the probability that the channel output is contained in a given bounded subset of $\mathbb{R}$ goes to zero.

and $\mathsf{K} = \{1, \ldots, K\}$ be index sets. The transpose, the determinant and the trace of a matrix $\boldsymbol{A}$ are denoted by $\boldsymbol{A}^\mathsf{T}$, $\det \boldsymbol{A}$, and $\mathrm{tr}(\boldsymbol{A})$, respectively. An identity matrix of size n is denoted by $\boldsymbol{I}_n$. The real part and imaginary part of a complex matrix $\boldsymbol{A}$ are denoted by $\mathrm{re}(\boldsymbol{A})$ and $\mathrm{im}(\boldsymbol{A})$, respectively. Let $\boldsymbol{j}$ be the imaginary number $\sqrt{-1}$. Then for any complex matrix $\boldsymbol{A}$, we have $\boldsymbol{A} = \mathrm{re}(\boldsymbol{A}) + \boldsymbol{j}\,\mathrm{im}(\boldsymbol{A})$.

For any $\boldsymbol{x} = [x_1, \ldots, x_N]^\mathsf{T} \in \mathbb{R}^N$, its amplitude $\|\boldsymbol{x}\|$ is defined as $\sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}}$ and we say its **phase**[1] is the normalized vector $\boldsymbol{x}/\|\boldsymbol{x}\|$. Let $\mathbb{B}_N(s)$ be the ball in $\mathbb{R}^N$ of radius $\sqrt{s}$, i.e.,

$$\mathbb{B}_N(s) = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}^\mathsf{T}\boldsymbol{x} \le s\}.$$

Then, the phase of a vector $\boldsymbol{x}$ is defined on the "surface" of the unit ball $\mathbb{B}_N(1)$. Clearly, $\boldsymbol{x}$ is uniquely determined by its amplitude and its phase. If $N = 1$, then $\boldsymbol{x}$ is a scalar and is simply denoted by x.

Let $\boldsymbol{X}$ be a random variable defined on $\mathbb{R}^N$ and μ be its probability measure. Then μ is called **discrete in amplitude and uniform in phase** (DAUP) [11] if the probability distribution of $\|\boldsymbol{X}\|$ is discrete with a finite number of probability mass points, and the phase of $\boldsymbol{X}$ is uniformly distributed on the surface of $\mathbb{B}_N(1)$.

B. Channel Model

Consider a discrete-time memoryless channel whose input is a real-valued N-tuple $\boldsymbol{X} = [X_1, \ldots X_N]^\mathsf{T}$ subject to input constraints to be defined and whose output is a real-valued M-tuple $\boldsymbol{Y} = [Y_1, \ldots Y_M]^\mathsf{T}$. In a practical communication system, due to various physical restrictions imposed on the system, not every channel input $\boldsymbol{x} \in \mathbb{R}^N$ can be generated by the transmitter. For example, all amplifiers have a limitation on their maximum allowable input and output amplitudes. A channel input $\boldsymbol{x}$ is said to be **admissible** if it can be produced by the transmitter. We will denote the set of admissible channel input vectors by $\mathbb{S}$, and we will assume that $\mathbb{S}$ is a closed and bounded subset of $\mathbb{R}^N$, unless specified explicitly. We refer to this channel input constraint as the **bounded-input constraint**. If the probability measure of $\boldsymbol{X}$ is μ, then the bounded-input constraint means that $\mu(\mathbb{S}) = 1$. For example, under a peak total power constraint in N dimensions of s^2, we have $\mathbb{S} = \mathbb{B}_N(s^2)$.

We call a channel **conditionally Gaussian (CG)** if the conditional probability distribution of the channel output is Gaussian distributed. In particular, for a given input $\boldsymbol{x} \in \mathbb{S}$, the conditional expectation vector and the conditional covariance matrix of the channel output $\boldsymbol{Y}$ are denoted by $\boldsymbol{b}_{\boldsymbol{x}}$ and $\Delta_{\boldsymbol{x}}$. Thus, the conditional probability density function $P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ is given by

$$P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{b}_{\boldsymbol{x}})^\mathsf{T} \nabla_{\boldsymbol{x}} (\boldsymbol{y} - \boldsymbol{b}_{\boldsymbol{x}}))}{\sqrt{(2\pi)^M \det \Delta_{\boldsymbol{x}}}} \tag{1}$$

where $\nabla_{\boldsymbol{x}}$ is the matrix inverse of $\Delta_{\boldsymbol{x}}$.

In this paper, the following assumptions are made: 1) for all $\boldsymbol{x} \in \mathbb{S}$, the covariance matrix $\Delta_{\boldsymbol{x}}$ is positive-definite (hence, $\det \Delta_{\boldsymbol{x}} > 0$), and 2) the entries in $\boldsymbol{b}_{\boldsymbol{x}}$ and $\Delta_{\boldsymbol{x}}$ are **well behaved**, in the sense that they can be extended holomorphically over $\mathbb{C}^N$ and are real over $\mathbb{R}^N$. For clarity, when the functions are extended over $\mathbb{C}^N$, a complex variable $\boldsymbol{w}$ is used instead of $\boldsymbol{x}$ to denote the function argument. Since $\mathbb{S}$ is assumed to be closed and bounded, the continuity of $\boldsymbol{b}_{\boldsymbol{x}}$ and $\Delta_{\boldsymbol{x}}$ implies that there exist $\kappa_l, \kappa_h, \vartheta > 0$ such that for all $\boldsymbol{x} \in \mathbb{S}$, all eigenvalues of $\nabla_{\boldsymbol{x}}$ are within the interval (κ_l, κ_h) and $\|\boldsymbol{b}_{\boldsymbol{x}}\| < \vartheta$. Hence, $1/\kappa_h^M \le \det \Delta_{\boldsymbol{x}} \le 1/\kappa_l^M$.

We also introduce cost constraints, as some channel inputs require more "effort" by the transmitter to generate than others. In this paper, we consider a set of K cost measures $g_k(\boldsymbol{x})$ indexed by K. The kth average cost constraint is characterized by the inequality $G_k(\mu) \le 0$, where $G_k(\mu) \triangleq \boldsymbol{E}_\mu[\, g_k(\boldsymbol{X})\,]$ and the expectation is taken with respect to the input probability measure μ. For instance, the commonly used constraint that the average signal power should not exceed γ is expressed

[1] To follow the similar terminology defined in previous work, the way we use the term "phase" in this paper is slightly different from the traditional usage in which it refers to the special case when $N = 2$ and the normalized vector is represented by an angle.

as $\boldsymbol{E}_\mu\left[\,||\boldsymbol{X}||^2-\gamma\,\right]\le 0$. Again, we assume that the cost functions $g_k(\boldsymbol{x})$ are all well behaved.

For simplicity, we use the tuple $(\Delta_{\boldsymbol{x}},\boldsymbol{b}_{\boldsymbol{x}},\{g_k(\boldsymbol{x}):k\in\mathsf{K}\},\mathbb{S})$ to denote a CG channel subject to the above bounded-input constraint and average cost constraints.

We call a CG channel **quadratic** if its input and output are related by the equation $\boldsymbol{Y}=\boldsymbol{H}\boldsymbol{X}+\boldsymbol{Z}$, where 1) $\boldsymbol{H}$ is a random $M\times N$ channel matrix whose entries are Gaussian distributed, and 2) $\boldsymbol{Z}=[Z_1,Z_2,\ldots,Z_M]^{\mathsf{T}}$ is an additive Gaussian noise that is independent of $\boldsymbol{H}$ and $\boldsymbol{X}$. For the special case that $\boldsymbol{H}$ is deterministic, we call the channel **constant quadratic**, or simply **constant**.

Clearly, for a quadratic CG channel, $\boldsymbol{b}_{\boldsymbol{x}}=\boldsymbol{E}[\boldsymbol{H}]\boldsymbol{x}+\boldsymbol{E}[\boldsymbol{Z}]$, and

$$\Delta_{\boldsymbol{x}}=\boldsymbol{E}[\boldsymbol{H}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}}]-\boldsymbol{E}[\boldsymbol{H}]\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{E}[\boldsymbol{H}^{\mathsf{T}}]+\boldsymbol{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}]-\boldsymbol{E}[\boldsymbol{Z}]\boldsymbol{E}[\boldsymbol{Z}^{\mathsf{T}}].$$

If the channel is constant, then $\boldsymbol{b}_{\boldsymbol{x}}=\boldsymbol{H}\boldsymbol{x}+\boldsymbol{E}[\boldsymbol{Z}]$ and

$$\Delta_{\boldsymbol{x}}=\boldsymbol{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}]-\boldsymbol{E}[\boldsymbol{Z}]\boldsymbol{E}[\boldsymbol{Z}^{\mathsf{T}}].$$

Thus, $\Delta_{\boldsymbol{x}}$ is constant for all $\boldsymbol{x}$ and will simply be denoted by Δ.

The following are some communication system channel models that fall within our framework.

1) A Scalar Additive White Gaussian Channel: Subject to peak and average power constraints, the channel is characterized by the tuple

$$(\sigma^2,x,\{x^2-\gamma\},\,[-s,s]).$$

The discreteness of the capacity-achieving measure for this channel was first shown in [1]. It is obvious that a scalar additive Gaussian channel is a constant CG channel whose channel "matrix" $\boldsymbol{H}$ is equal to 1.

2) A System of N Parallel Independent Gaussian Channels: Subject to peak and average total power constraints, the channel is characterized by the tuple

$$(\sigma^2\boldsymbol{I}_N,\boldsymbol{x},\{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}-\gamma\},\mathbb{B}_N(s^2)).$$

This is a generalization of the scalar-Gaussian channel. The special case when $N=2$ was studied in [3], and it was shown that the capacity-achieving measure is DAUP.

3) An Optical Intensity Modulated/Direct Detected Channel (Using Raised Quadrature Amplitude Modulation (QAM) Basis Functions [14]): Under the assumption that the noise is dominated by the background illumination, subject to peak and average power constraints, the channel is a constant CG channel and is characterized by the tuple

$$(\sigma^2\boldsymbol{I}_3,\boldsymbol{x},\{x_1-\gamma\},\mathbb{S})$$

where

$$\mathbb{S}=\left\{\boldsymbol{x}\in\mathbb{R}^3:\ \sqrt{2x_2^2+2x_3^2}\le x_1\le s-\sqrt{2x_2^2+2x_3^2}\right\}. \tag{2}$$

Here, for any input $\boldsymbol{x}$, the optical power is given by x_1. See [14] for more details concerning this channel model.

4) A Signal-Dependent Noise Optical Pulse Amplitude Modulation (PAM) Channel: Subject to peak and average power constraints, the channel is characterized by the tuple

$$(\sigma_1^2x+\sigma_0^2,x,\{x-\gamma\},[0,s]).$$

This example will be elaborated in detail in Section IV-A.

5) A MIMO Rayleigh-Fading Channel: Assuming that there are $N/2$ and $M/2$ transmitting and receiving antennae, subject to peak and average power constraints, the channel is characterized by the tuple

$$\left(\frac{1}{2}(\chi^2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}+\sigma^2)\boldsymbol{I}_M,\boldsymbol{0}\,,\{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}-\gamma\},\mathbb{B}_N(s^2)\right).$$

It was proved in [4] that the capacity-achieving measure for this channel is discrete in the sense that the distribution of $||\boldsymbol{X}||$ is discrete with a finite number of probability mass points. Furthermore, the discreteness of the capacity-achieving measure remains valid even when $s=\infty$, i.e., $\mathbb{S}=\mathbb{R}^N$.

6) A MIMO Rician Fading Channel: Subject to a peak power constraint, the channel is characterized by the tuple

$$\left(\frac{1}{2}(\chi^2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}+\sigma^2)\boldsymbol{I}_M,\boldsymbol{E}[\boldsymbol{H}]\boldsymbol{x},\{\},\mathbb{B}_N(s^2)\right).$$

7) A MIMO Rayleigh Fading Channel With Receiver-Side Channel Information: Assuming that there are $N/2$ transmitting and one receiving antennae, subject to peak power constraint, the channel is characterized by the tuple

$$(\Delta_{\boldsymbol{x}},\boldsymbol{0},\{\},\mathbb{B}_N(s^2))$$

where $\boldsymbol{x}^{\mathsf{T}}=[\boldsymbol{x}_{\mathrm{re}}^{\mathsf{T}},\boldsymbol{x}_{\mathrm{im}}^{\mathsf{T}}]$ and $\Delta_{\boldsymbol{x}}$ is defined as

$$\frac{1}{2}\begin{bmatrix}\chi^2\boldsymbol{I}_N & \boldsymbol{0} & \chi^2\boldsymbol{x}_{\mathrm{re}} & \chi^2\boldsymbol{x}_{\mathrm{im}}\\ \boldsymbol{0} & \chi^2\boldsymbol{I}_N & -\chi^2\boldsymbol{x}_{\mathrm{im}} & \chi^2\boldsymbol{x}_{\mathrm{re}}\\ \chi^2\boldsymbol{x}_{\mathrm{re}}^{\mathsf{T}} & -\chi^2\boldsymbol{x}_{\mathrm{im}}^{\mathsf{T}} & \chi^2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}+\sigma^2 & 0\\ \chi^2\boldsymbol{x}_{\mathrm{im}}^{\mathsf{T}} & \chi^2\boldsymbol{x}_{\mathrm{re}}^{\mathsf{T}} & 0 & \chi^2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}+\sigma^2\end{bmatrix}. \tag{3}$$

8) An Interference Channel: Subject to peak power constraint, the channel is characterized by the tuple

$$((\chi^2\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}+\sigma^2)\boldsymbol{I}_N\ ,\ \boldsymbol{x}\ ,\ \{\}\ ,\ \mathbb{B}_N(s^2)).$$

This channel is a special case of MIMO Rician fading channel where $\boldsymbol{E}[\boldsymbol{H}]$ is equal to $\boldsymbol{I}_N$. We will elaborate more in Section IV-C.

III. Main Results

Consider a CG channel characterized by the tuple $(\Delta_{\boldsymbol{x}},\boldsymbol{b}_{\boldsymbol{x}},\{g_k(\boldsymbol{x}):k\in\mathsf{K}\},\mathbb{S})$. We are interested in the probability measure which satisfies the channel input constraints and maximizes the mutual information between the channel input and output. Let $\Lambda_{\mathbb{S}}$ be the set of all input probability measures μ satisfying the bounded-input constraint, and Λ be the subset of $\Lambda_{\mathbb{S}}$ which also satisfies the average cost constraints (if any). For any input probability measure μ of $\boldsymbol{X}$, we denote the μ-induced probability density function of $\boldsymbol{Y}$ by $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)$, the differential entropy of $\boldsymbol{Y}$ by $H_{\boldsymbol{Y}}(\mu)$, the conditional differential entropy of

$\boldsymbol{Y}$ given $\boldsymbol{X}$ by $H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu)$, and the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by $I(\mu)$. Thus,

$$\begin{aligned} P_{\boldsymbol{Y}}(\boldsymbol{y};\mu) &= \int P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})\, d\mu \\ H_{\boldsymbol{Y}}(\mu) &= -\int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)\, d\boldsymbol{y} \\ H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu) &= \frac{1}{2}\int \log[(2\pi e)^M \det \Delta_{\boldsymbol{x}}]\, d\mu \\ I(\mu) &= H_{\boldsymbol{Y}}(\mu) - H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu). \end{aligned} \tag{4}$$

As a result, the **channel capacity problem** is the following optimization problem:

$$\begin{cases} \text{Maximize} & I(\mu) \\ \text{subject to} & \mu \in \Lambda_{\mathbb{S}}, \text{ and for all } k \in \mathsf{K},\ G_k(\mu) \le 0. \end{cases} \tag{5}$$

Our first result shows that a capacity-achieving measure exists.

Theorem 1 (Existence and Uniqueness): There exists μ_o in Λ which solves the channel capacity problem. Furthermore, the capacity-achieving output distribution is unique. In other words, if ν is also a capacity-achieving input measure, then $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o) = P_{\boldsymbol{Y}}(\boldsymbol{y};\nu)$.

Suppose, in addition, that the CG channel is constant and the channel matrix $\boldsymbol{H}$ is left-invertible. Then the capacity-achieving input measure is unique.

Sketch of Proof: The proof of Theorem 1 follows the same approach used in [1]. First, we will prove that Λ is convex and sequentially compact[2] (see Proposition 2). As the mutual information function $I(\mu)$ is continuous over Λ (see Proposition 3), it achieves its maximum in Λ.

For any $0 < \theta < 1$ and $\mu_0, \mu_1 \in \Lambda_{\mathbb{S}}$, let $\mu_\theta = (1-\theta)\mu_0 + \theta\mu_1$. It is well known that $I(\mu)$ and $H_{\boldsymbol{Y}}(\mu)$ are concave, i.e., $I(\mu_\theta) \ge (1-\theta)I(\mu_0) + \theta I(\mu_1)$. By the strict concavity of the function $f(c) = -c\log c$, we can show that equality holds if and only if $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0) = P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)$.

Furthermore, if the CG channel is constant with a left-invertible channel matrix $\boldsymbol{H}$, then by Lemma 4, there is a one-to-one correspondence between the input and output probability measures, i.e., $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0) = P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)$ if and only if $\mu_0 = \mu_1$. As a result, $I(\mu)$ is a strictly concave function of μ. The uniqueness of the capacity-achieving input measure then follows. □

Definition 1: A point $\boldsymbol{x} \in \mathbb{R}^N$ is said to be a **point of increase** of μ if for any open subset $\mathbb{O}$ of $\mathbb{R}^N$ containing $\boldsymbol{x}$, $\mu(\mathbb{O}) > 0$.

Let $\mathbb{E}_\mu$ be the set of points of increase of μ. Then $\mu(\mathbb{E}_\mu) = 1$. In fact, $\mathbb{E}_\mu$ is the minimal closed subset of $\mathbb{R}^N$ whose probability is 1.

Theorem 2 (Necessity and Sufficiency): Let μ_o be an admissible input probability measure, i.e., $\mu_o \in \Lambda$. Let $C(\boldsymbol{x})$ be $\frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{x}}]$. Then μ_o is capacity achieving if and only if there exists $\{\lambda_k \ge 0 :\ k \in \mathsf{K}\}$ such that for all $\boldsymbol{x} \in \mathbb{S}$

$$Q(\boldsymbol{x};\mu_o) - C(\boldsymbol{x}) - I(\mu_o) - \sum_{k\in\mathsf{K}} \lambda_k g_k(\boldsymbol{x}) \le 0 \tag{6}$$

[2]A space is sequentially compact if every infinite sequence in the space has a convergent subsequence.

where $Q(\boldsymbol{x};\mu_o)$ is defined as $-\int P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o)\, d\boldsymbol{y}$. Furthermore, if $\boldsymbol{x}$ belongs to $\mathbb{E}_{\mu_o}$, the inequality (6) is satisfied with equality.

Sketch of Proof: The proof of this theorem also follows a similar approach as in [1]. By the method of Lagrange multipliers [16], μ_o is capacity achieving if and only if 1) there exists $\{\lambda_k \ge 0 :\ k \in \mathsf{K}\}$ such that $\lambda_k G_k(\mu_o) = 0$ for all $k \in \mathsf{K}$ and 2) for all $\mu \in \Lambda_{\mathbb{S}}$, $J(\mu_o) \ge J(\mu)$ where $J(\mu) = I(\mu) - \sum_{k\in\mathsf{K}} \lambda_k G_k(\mu)$. Since the function $J(\mu)$ is concave, the condition 2) is equivalent to that $J'_{\mu_o}(\mu) \le 0$ for all $\mu \in \Lambda_{\mathbb{S}}$ where $J'_{\mu_o}(\mu)$ is the weak derivative of $J(\mu)$ at μ_o (see Definition 6). Since $G_k(\mu)$ is a linear function of μ, we have

$$J'_{\mu_o}(\mu) = I'_{\mu_o}(\mu) - \sum_{k\in\mathsf{K}} \lambda_k (G_k(\mu) - G_k(\mu_o)).$$

Therefore, given that the condition 1) is true, the condition 2) is equivalent to

$$I'_{\mu_o}(\mu) - \sum_{k\in\mathsf{K}} \lambda_k G_k(\mu) \le 0 \tag{7}$$

for all $\mu \in \Lambda_{\mathbb{S}}$. In Proposition 4, we prove that

$$I'_{\mu_o}(\mu) = \int Q(\boldsymbol{x};\mu_o)\, d\mu - H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu) - I(\mu_o).$$

For simplicity, the function $a(\boldsymbol{x}, \mu_o, \{\lambda_k : k \in \mathsf{K}\})$ is used to denote the following formula:

$$Q(\boldsymbol{x};\mu_o) - \frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{x}}] - I(\mu_o) - \sum_{k\in\mathsf{K}} \lambda_k g_k(\boldsymbol{x}). \tag{8}$$

(Proof of Necessity) Suppose μ_o is capacity achieving. Let $\{\lambda_k \ge 0 : k \in \mathsf{K}\}$ be selected such that the inequality (7) is satisfied. For any $\boldsymbol{x}^* \in \mathbb{S}$, let μ be the probability measure such that $\mu(\{\boldsymbol{x}^*\}) = 1$, i.e., μ is the measure with a single probability mass at $\boldsymbol{x}^*$. Obviously, $\mu \in \Lambda_{\mathbb{S}}$. Substituting μ into (7), we have $a(\boldsymbol{x}^*, \mu_o, \{\lambda_k : k \in \mathsf{K}\}) \le 0$. The necessity is thus proved.

(Proof of Sufficiency) Suppose the inequality (6) is satisfied for selected $\{\lambda_k \ge 0 :\ k \in \mathsf{K}\}$. Since $\int Q(\boldsymbol{x};\mu_o)\, d\mu_o = H_{\boldsymbol{Y}}(\mu_o)$ and $\int C(\boldsymbol{x})\, d\mu_o = H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu_o)$, if we integrate both sides of (6) with respect to μ_o, we have

$$\begin{aligned} 0 &\ge \int Q(\boldsymbol{x};\mu_o)\, d\mu_o - \int C(\boldsymbol{x})\, d\mu_o - I(\mu_o) \\ &\quad - \sum_{k\in\mathsf{K}} \lambda_k \int g_k(\boldsymbol{x})\, d\mu_o \\ &= -\sum_{k\in\mathsf{K}} \lambda_k G_k(\mu_o) \\ &\ge 0 \end{aligned}$$

where the last inequality follows from that $G_k(\mu_o) \le 0$. Hence, $\lambda_k G_k(\mu_o) = 0$ for all $k \in \mathsf{K}$. On the other hand, if we integrate (6) with respect to any $\mu \in \Lambda_{\mathbb{S}}$, we have

$$\begin{aligned} 0 &\ge \int Q(\boldsymbol{x};\mu_o)\, d\mu - \int C(\boldsymbol{x})\, d\mu - I(\mu_o) \\ &\quad - \sum_{k\in\mathsf{K}} \lambda_k \int g_k(\boldsymbol{x})\, d\mu \qquad (9) \\ &= I'_{\mu_o}(\mu) - \sum_{k\in\mathsf{K}} \lambda_k G_k(\mu). \qquad (10) \end{aligned}$$

Hence, μ_o is capacity-achieving.

Finally, it remains to prove that the inequality (6) is satisfied with equality at $\mathbb{E}_{\mu_o}$. Let $\boldsymbol{x}^* \in \mathbb{E}_{\mu_o}$, and suppose, to the contrary, that $a(\boldsymbol{x}^*, \mu_o, \{\lambda_k : k \in \mathsf{K}\}) = -\epsilon < 0$. By the continuity of $Q(\boldsymbol{x}; \mu_o)$ (see Lemma 5), the function $a(\boldsymbol{x}, \mu_o, \{\lambda_k : k \in \mathsf{K}\})$ is also continuous on $\mathbb{S}$. Hence, there is an open subset $\mathbb{O}$ containing $\boldsymbol{x}^*$ such that $a(\boldsymbol{x}, \mu_o, \{\lambda_k : k \in \mathsf{K}\}) \leq -\epsilon/2$ for all $\boldsymbol{x} \in \mathbb{O}$. Consequently, if we integrate both sides of (6) with respect to μ_o, we have $-\epsilon\mu_o(\mathbb{O})/2 \geq -\sum_{k\in\mathsf{K}} \lambda_k G_k(\mu_o) = 0$. A contradiction occurs, and the theorem is proved. □

By Theorem 1, if μ_o and ν are both capacity achieving, then $P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_o) = P_{\boldsymbol{Y}}(\boldsymbol{y}; \nu)$, which further implies that $Q(\boldsymbol{x}; \mu_o) = Q(\boldsymbol{x}; \nu)$. Hence,

$$a(\boldsymbol{x}, \mu_o, \{\lambda_k : k \in \mathsf{K}\}) = a(\boldsymbol{x}, \nu, \{\lambda_k : k \in \mathsf{K}\})$$

for all $\boldsymbol{x} \in \mathbb{S}$. According to Theorem 2, the inequality (6) is thus tight at any points of increase of a capacity-achieving measure.

Definition 2: A complex-valued function $f(\boldsymbol{w})$ defined on an open subset $\mathbb{U}$ of $\mathbb{C}^N$ is called **holomorphic** [17] on $\mathbb{U}$ if it is analytic in each individual variable on $\mathbb{U}$.

Definition 3: A subset $\mathbb{F}$ of $\mathbb{R}^N$ is defined to be **sparse** (in $\mathbb{R}^N$) if there exists a nonzero holomorphic function f defined on a connected open subset $\mathbb{U}$ of $\mathbb{C}^N$ containing the closure of $\mathbb{F}$ such that $f(\boldsymbol{w}) = 0$ for all $\boldsymbol{w} \in \mathbb{F}$.

Suppose $\mathbb{F}$ is not sparse and $f(\boldsymbol{w})$ is holomorphic on a connected open subset $\mathbb{U}$ of $\mathbb{C}^N$ containing the closure of $\mathbb{F}$. If f is zero on $\mathbb{F}$, then it is also zero on $\mathbb{U}$.

By the identity theorem, a bounded subset $\mathbb{F}$ of $\mathbb{R}$ is sparse in $\mathbb{R}$ if and only if it is finite. However, for a high-dimensional space $\mathbb{R}^N$, it is difficult in general to determine whether a bounded subset of $\mathbb{R}^N$ is sparse or not. In the following lemma, we show that when $\mathbb{F}$ is a collection of "concentric shells" and is bounded, it is sparse if and only if its number of shells is finite.

Lemma 1: Suppose $\{p_i : i \in \mathsf{I}\}$ is a set of distinct nonnegative real numbers which are bounded above by a positive real number. Let $\mathbb{F}$ be a collection of "concentric shells" of radii $\{p_i : i \in \mathsf{I}\}$, i.e.,

$$\mathbb{F} = \bigcup_{i\in\mathsf{I}} \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} = p_i^2\}.$$

Then $\mathbb{F}$ is sparse if and only if I is finite.

Proof: See Appendix A. □

Definition 4: A probability measure μ is said to be **discrete** if its set of points of increase $\mathbb{E}_\mu$ is sparse.

Definition 5: Let $\Delta_{\boldsymbol{w}}$ be a holomorphic extension of $\Delta_{\boldsymbol{x}}$ to $\mathbb{C}^N$. The **well-behaved region** of $\Delta_{\boldsymbol{w}}$, denoted by $\mathbb{A}(\Delta_{\boldsymbol{w}})$, is the subset of $\mathbb{C}^N$ such that for all its elements $\boldsymbol{w}$, $\mathrm{re}(\det \Delta_{\boldsymbol{w}}) > 0$ and $\mathrm{re}(\nabla_{\boldsymbol{w}})$ is positive definite.

Note that $\mathrm{re}(\det \Delta_{\boldsymbol{w}}) > 0$ implies that the logarithm and the square root of $\det \Delta_{\boldsymbol{w}}$ are well defined and holomorphic over $\mathbb{A}(\Delta_{\boldsymbol{w}})$. Furthermore, as $\mathrm{re}(\nabla_{\boldsymbol{w}})$ is a symmetric matrix, the condition that $\mathrm{re}(\nabla_{\boldsymbol{w}})$ is positive definite is equivalent to the condition that the eigenvalues of $\mathrm{re}(\nabla_{\boldsymbol{w}})$ are positive. As the eigenvalues of $\mathrm{re}(\nabla_{\boldsymbol{w}})$ are continuous functions of $\boldsymbol{w}$, the well-behaved region $\mathbb{A}(\Delta_{\boldsymbol{w}})$ is an open subset in $\mathbb{C}^N$. In the special case when the channel is constant, $\Delta_{\boldsymbol{w}}$ is real, constant, and positive definite for all $\boldsymbol{w} \in \mathbb{C}^N$. Hence, $\mathbb{A}(\Delta) = \mathbb{C}^N$.

Theorem 3 (Discreteness): Suppose μ_o is capacity achieving and $\{\lambda_k \geq 0 : k \in \mathsf{K}\}$ are chosen such that the inequality (6) is satisfied. Let $\mathbb{E}$ be the subset of $\mathbb{S}$ at which the inequality (6) is tight. In the following three cases, the set $\mathbb{E}$ is sparse, and hence, μ_o is discrete.

- [Case A] There is no average cost constraint (i.e., K is empty), and there exists a connected open subset $\mathbb{W}$ of $\mathbb{A}(\Delta_{\boldsymbol{w}})$ containing $\mathbb{S}$ and a sequence $\{\boldsymbol{w}^{(i)}\}_{i=1}^{\infty}$ in $\mathbb{W}$ such that 1) $\boldsymbol{b}_{\boldsymbol{w}^{(i)}}$ and $\Delta_{\boldsymbol{w}^{(i)}}$ are real for all positive integers i, and 2) $\overline{\lim}_{i\to\infty} \mathrm{tr}(\Delta_{\boldsymbol{w}^{(i)}}) = +\infty$.
- [Case B] There exists a connected open subset $\mathbb{W}$ of $\mathbb{A}(\Delta_{\boldsymbol{w}})$ containing $\mathbb{S}$ and a convergent sequence $\{\boldsymbol{w}^{(i)}\}_{i=1}^{\infty}$ in $\mathbb{W}$ with a limit $\boldsymbol{w}^{(0)}$ (not necessary in $\mathbb{W}$), such that 1) $\boldsymbol{b}_{\boldsymbol{w}^{(i)}}$ and $\Delta_{\boldsymbol{w}^{(i)}}$ are real for all positive integers i, and 2) $\lim_{i\to\infty} \det \Delta_{\boldsymbol{w}^{(i)}} = 0$.
- [Case C] The channel is constant with a nonzero channel matrix $\boldsymbol{H}$, and the cost functions are of second order, i.e., $g_k(\boldsymbol{x}) \triangleq \boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}_k\boldsymbol{x} + \boldsymbol{L}_k\boldsymbol{x} - \gamma_k$ for some $N \times N$ real matrices $\boldsymbol{Q}_k$, $1 \times N$ real row vectors $\boldsymbol{L}_k$ and real scalars γ_k.

Sketch of Proof: Let μ_o be the capacity-achieving measure for the channel $(\Delta_{\boldsymbol{x}}, \boldsymbol{b}_{\boldsymbol{x}}, \{g_k(\boldsymbol{x}) : k \in \mathsf{K}\}, \mathbb{S})$. For any $\boldsymbol{w} \in \mathbb{A}(\Delta_{\boldsymbol{w}})$, define $d(\boldsymbol{w})$ as

$$-\int \frac{\exp(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})^{\mathsf{T}}\nabla_{\boldsymbol{w}}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}}))}{\sqrt{(2\pi)^M \det \Delta_{\boldsymbol{w}}}} \log P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_o)\, d\boldsymbol{y}. \tag{11}$$

It is clear from the definition that for all $\boldsymbol{x} \in \mathbb{S}$, $d(\boldsymbol{x}) = Q(\boldsymbol{x}; \mu_o)$. Let

$$a(\boldsymbol{w}) \triangleq d(\boldsymbol{w}) - \frac{1}{2}\log(2\pi e)^M \det \Delta_{\boldsymbol{w}} - I(\mu_o) - \sum_{k\in\mathsf{K}} \lambda_k g_k(\boldsymbol{w}).$$

Then, by definition, $a(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{E}$.

In the above three cases, there exists a connected open subset $\mathbb{W}$ of $\mathbb{A}(\Delta_{\boldsymbol{w}})$ containing $\mathbb{S}$. In particular, $\mathbb{W}$ is equal to $\mathbb{C}^N$ in Case C. Since $\mathbb{S}$ is a closed and bounded subset containing $\mathbb{E}$, $\mathbb{W}$ also contains the closure of $\mathbb{E}$. As $d(\boldsymbol{w})$ is holomorphic on $\mathbb{A}(\Delta_{\boldsymbol{w}})$ (see Proposition 5), $a(\boldsymbol{w})$ is also holomorphic on $\mathbb{A}(\Delta_{\boldsymbol{w}})$. Hence, if $\mathbb{E}$ is not sparse, then for all $\boldsymbol{w} \in \mathbb{W}$, $a(\boldsymbol{w}) = 0$, or equivalently

$$d(\boldsymbol{w}) = \frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{w}}] + I(\mu_o) + \sum_{k\in\mathsf{K}} \lambda_k g_k(\boldsymbol{w}).$$

In the following, we will show that if $\mathbb{E}$ is not sparse, then a contradiction occurs in each of the three cases and hence Theorem 3 is proved.

Contradiction in Case A:

Suppose the requirements in Case A are satisfied, and $\mathbb{E}$ is not sparse. By Lemma 6, there exist $\alpha, \beta > 0$ such that $d(\boldsymbol{w}^{(i)}) \geq -\alpha + \beta \mathrm{tr}[\Delta_{\boldsymbol{w}^{(i)}}]$. Without loss of generality, assume that $\xi_1(\boldsymbol{w}^{(i)}) \geq \cdots \geq \xi_M(\boldsymbol{w}^{(i)}) > 0$ are the eigenvalues of $\Delta_{\boldsymbol{w}^{(i)}}$. Then

$$\mathrm{tr}[\Delta_{\boldsymbol{w}^{(i)}}] = \sum_{m\in\mathsf{M}} \xi_m(\boldsymbol{w}^{(i)})$$

and

$$\det \Delta_{\boldsymbol{w}^{(i)}} = \prod_{m\in\mathsf{M}} \xi_m(\boldsymbol{w}^{(i)}).$$

As a result

$$\frac{1}{2}\log(2\pi e)^M + \frac{1}{2}\sum_{m\in\mathsf{M}} \log \xi_m(\boldsymbol{w}^{(i)}) + I(\mu_o) \geq -\alpha + \beta \sum_{m\in\mathsf{M}} \xi_m(\boldsymbol{w}^{(i)})$$

and hence,

$$\frac{1}{2}\log(2\pi e)^M + I(\mu_o) + \alpha \geq \beta\xi_1(\boldsymbol{w}^{(i)}) - \frac{M}{2}\log \xi_1(\boldsymbol{w}^{(i)}).$$

Since

$$\overline{\lim_{i\to\infty}} \operatorname{tr}(\Delta_{\boldsymbol{w}^{(i)}}) = +\infty$$

$\xi_1(\boldsymbol{w}^{(i)})$ and $\beta\xi_1(\boldsymbol{w}^{(i)}) - \frac{M}{2}\log \xi_1(\boldsymbol{w}^{(i)})$ are not bounded above. A contradiction thus occurs.

Contradiction in Case B:

Suppose the requirements in Case B are satisfied, and $\mathbb{E}$ is not sparse. Again, by Lemma 6, there exist $\alpha, \beta > 0$ such that $d(\boldsymbol{w}^{(i)}) \geq -\alpha + \beta \operatorname{tr}[\Delta_{\boldsymbol{w}^{(i)}}]$. Therefore,

$$\frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{w}^{(i)}}] \geq -\alpha + \beta \operatorname{tr}[\Delta_{\boldsymbol{w}^{(i)}}] - I(\mu_o) - \sum_{k\in\mathsf{K}} \lambda_k g_k(\boldsymbol{w}^{(i)})$$

and consequently

$$\lim_{i\to\infty} \frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{w}^{(i)}}] \geq -\alpha + \beta \operatorname{tr}[\Delta_{\boldsymbol{w}^{(0)}}] - I(\mu_o) - \sum_{k\in\mathsf{K}} \lambda_k g_k(\boldsymbol{w}^{(0)}).$$

Since $\lim_{i\to\infty} \log\det \Delta_{\boldsymbol{w}^{(i)}} = -\infty$, a contradiction thus occurs.

Contradiction in Case C:

Suppose the requirements in Case C are satisfied, and $\mathbb{E}$ is not sparse. Then, for all $\boldsymbol{x} \in \mathbb{R}^N$, we have

$$d(\boldsymbol{x}) = H_{\boldsymbol{Y}}(\mu_o) + \sum_{k\in\mathsf{K}} \lambda_k(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}_k\boldsymbol{x} + \boldsymbol{L}_k\boldsymbol{x} - \gamma_k). \tag{12}$$

Following the same approach as in [1], we will show that the equality (12) implies that the capacity-achieving output distribution is Gaussian. Since $\boldsymbol{Y} = \boldsymbol{HX} + \boldsymbol{Z}$ and $\boldsymbol{Z}$ is Gaussian distributed, $\boldsymbol{HX}$ is also Gaussian distributed. As the set of points of increase of a Gaussian distribution is not bounded, the Gaussianity of $\boldsymbol{HX}$ contradicts the assumption that μ_o satisfies the bounded-input constraint.

Now, it remains to prove that $\boldsymbol{Y}$ is Gaussian distributed. First, assume that the rank of $\boldsymbol{H}$ is equal to M. Without loss of generality, we can also assume that $\boldsymbol{Z}$ is zero mean. Let $P_{\boldsymbol{Z}}(\boldsymbol{z})$ be the probability density function of the additive Gaussian noise. Then, it is clear that for $\boldsymbol{x} \in \mathbb{R}^N$

$$d(\boldsymbol{x}) = -\int P_{\boldsymbol{Z}}(\boldsymbol{Hx} - \boldsymbol{y}) \log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o)\, d\boldsymbol{Y} \tag{13}$$

$$= -P_{\boldsymbol{Z}}(\boldsymbol{Hx}) * \log P_{\boldsymbol{Y}}(\boldsymbol{Hx};\mu_o) \tag{14}$$

where $*$ denotes convolution. Let $\boldsymbol{r} = \boldsymbol{Hx}$. The right-hand side of (14) depends only on $\boldsymbol{r}$. Hence,

$$H_{\boldsymbol{Y}}(\mu_o) + \sum_{k\in\mathsf{K}} \lambda_k(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}_k\boldsymbol{x} + \boldsymbol{L}_k\boldsymbol{x} - \gamma_k)$$

also depends on $\boldsymbol{r}$ only and we can thus construct matrices $\widehat{\boldsymbol{Q}}_k$ and $\widehat{\boldsymbol{L}}_k$ such that

$$H_{\boldsymbol{Y}}(\mu_o) + \sum_{k\in\mathsf{K}} \lambda_k(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}_k\boldsymbol{x} + \boldsymbol{L}_k\boldsymbol{x} - \gamma_k) = H_{\boldsymbol{Y}}(\mu_o) + \sum_{k\in\mathsf{K}} \lambda_k(\boldsymbol{r}^{\mathsf{T}}\widehat{\boldsymbol{Q}}_k\boldsymbol{r} + \widehat{\boldsymbol{L}}_k\boldsymbol{r} - \gamma_k).$$

Thus, we have

$$-P_{\boldsymbol{Z}}(\boldsymbol{r}) * \log P_{\boldsymbol{Y}}(\boldsymbol{r};\mu_o) = H_{\boldsymbol{Y}}(\mu_o) + \sum_{k\in\mathsf{K}} \lambda_k(\boldsymbol{r}^{\mathsf{T}}\widehat{\boldsymbol{Q}}_k\boldsymbol{r} + \widehat{\boldsymbol{L}}_k\boldsymbol{r} - \gamma_k). \tag{15}$$

By direct integration, we can immediately verify that $\widehat{\boldsymbol{L}}_k\boldsymbol{r} = P_{\boldsymbol{Z}}(\boldsymbol{r}) * \widehat{\boldsymbol{L}}_k\boldsymbol{r}$ and

$$\boldsymbol{r}^{\mathsf{T}}\widehat{\boldsymbol{Q}}_k\boldsymbol{r} = P_{\boldsymbol{Z}}(\boldsymbol{r}) * \left(\boldsymbol{r}^{\mathsf{T}}\widehat{\boldsymbol{Q}}_k\boldsymbol{r} - \sum_{i,j\in\mathsf{M}} \widehat{\boldsymbol{Q}}_k^{(i,j)}\Delta^{(i,j)}\right)$$

where $\widehat{\boldsymbol{Q}}_k^{(i,j)}$ and $\Delta^{(i,j)}$ are the (i,j)th entries of $\widehat{\boldsymbol{Q}}_k$ and Δ, respectively. Hence, (15) implies that for all $\boldsymbol{r} \in \mathbb{R}^M$, we have

$$P_{\boldsymbol{Z}}(\boldsymbol{r}) * \left(\begin{array}{l}\log P_{\boldsymbol{Y}}(\boldsymbol{r};\mu_o) + H_{\boldsymbol{Y}}(\mu_o) \\ + \sum_{k\in\mathsf{K}} \lambda_k \left(\begin{array}{l}\boldsymbol{r}^{\mathsf{T}}\widehat{\boldsymbol{Q}}_k\boldsymbol{r} + \widehat{\boldsymbol{L}}_k\boldsymbol{r} - \gamma_k \\ - \sum_{i,j\in\mathsf{M}} \widehat{\boldsymbol{Q}}_k^{(i,j)}\Delta^{(i,j)}\end{array}\right)\end{array}\right) = \boldsymbol{0}.$$

As the Fourier transform of $P_{\boldsymbol{Z}}$ is nonzero everywhere, by Corollary 9, the function

$$\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o) + H_{\boldsymbol{Y}}(\mu_o) + \sum_{k\in\mathsf{K}} \lambda_k \left(\boldsymbol{y}^{\mathsf{T}}\widehat{\boldsymbol{Q}}_k\boldsymbol{y} - \sum_{i,j\in\mathsf{M}} \widehat{\boldsymbol{Q}}_k^{(i,j)}\Delta^{(i,j)} + \widehat{\boldsymbol{L}}_k\boldsymbol{y} - \gamma_k\right)$$

is equal to 0, which further implies that $\boldsymbol{Y}$ is Gaussian distributed.

Now, assume that the rank of $\boldsymbol{H}$ is equal to $M' < M$. As Δ is real, symmetric, and positive definite, there exists an invertible matrix $\boldsymbol{B}$ such that $\boldsymbol{B}\Delta\boldsymbol{B}^{\mathsf{T}} = \boldsymbol{I}_M$. Using singular-value decomposition on $\boldsymbol{BH}$, we can construct an $M' \times M$ matrix $\boldsymbol{V}$ such that 1) $I(\boldsymbol{X};\boldsymbol{HX} + \boldsymbol{Z}) = I(\boldsymbol{X};\boldsymbol{VBHX} + \boldsymbol{VBZ})$, and 2) the rank of $\boldsymbol{VBH}$ is M'. Hence, the original constant CG channel is equivalent to another constant CG channel where the channel matrix is $\boldsymbol{VBH}$ and the additive noise is $\boldsymbol{VBZ}$. Using the same argument as above, we can conclude that if the requirements in Case C are satisfied, then $\mathbb{E}$ must be sparse. □

As a remark, subject to an average power constraint ($\boldsymbol{E}_\mu[\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma] \leq 0$), the capacity-achieving measure for a constant CG channel whose channel gain matrix is an identity matrix has also been studied separately in [8], in which it was proved that the capacity-achieving measure is discrete if $\mathbb{R}^N \backslash \mathbb{S}$ has a positive Lebesgue measure.

Corollary 1: Let $(\Delta_{\boldsymbol{x}}, \boldsymbol{b}_{\boldsymbol{x}}, \{\}, \mathbb{S})$ be a quadratic CG channel subject only to a bounded-input constraint. The ca-

pacity-achieving measure is discrete except in the trivial case that the channel is constant with a zero channel matrix $\boldsymbol{H}$.

Proof: Let $\boldsymbol{H}$ be the channel matrix for the quadratic CG channel. Then

$$\Delta_{\boldsymbol{x}} = \boldsymbol{E}[\boldsymbol{H}\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{H}^{\mathsf{T}}] - \boldsymbol{E}[\boldsymbol{H}]\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}\boldsymbol{E}[\boldsymbol{H}^{\mathsf{T}}] + \boldsymbol{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}] - \boldsymbol{E}[\boldsymbol{Z}]\boldsymbol{E}[\boldsymbol{Z}^{\mathsf{T}}].$$

Clearly, $\Delta_{\boldsymbol{x}}$ and $\boldsymbol{b}_{\boldsymbol{x}}$ are real over $\mathbb{R}^N$, and $\mathbb{R}^N$ is a subset of $\mathbb{A}(\Delta_{\boldsymbol{w}})$.

Suppose the quadratic CG channel is not constant. Let $\mathbb{W}$ be an open subset of $\mathbb{A}(\Delta_{\boldsymbol{w}})$ containing $\mathbb{R}^N$. Then $\mathrm{tr}[\Delta_{\boldsymbol{x}}]$ is unbounded over $\mathbb{W}$. Hence, the requirements in Case A of Theorem 3 are satisfied, and thus the capacity-achieving measure is discrete.

If the quadratic CG channel is constant, then the requirements in Case C of Theorem 3 are also satisfied. Hence, the capacity-achieving measure is still discrete. □

It is well known that for a scalar additive Gaussian channel subject only to an average power constraint, the capacity-achieving input distribution is Gaussian. When an additional bounded-input constraint $\mathbb{S} = [-s, s]$ is imposed, the capacity-achieving measure becomes discrete. An interesting question then arises: if we relax the bounded-input constraint by increasing s, then what is its effect on the capacity-achieving measure? In the next theorem, we will prove that as s increases, the capacity-achieving measure approaches the capacity-achieving distribution when there is no bounded-input constraint.

Consider a sequence of CG channels

$$\{(\Delta, \boldsymbol{H}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{S}_j)\}_{j=1}^{\infty}$$

where $\boldsymbol{H}$ is a nonzero real matrix and $\mathbb{S}_j$ is a closed and bounded subset of $\mathbb{R}^N$. Let μ_j and Φ be the capacity-achieving measures for the channels $(\Delta, \boldsymbol{H}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{S}_j)$ and $(\Delta, \boldsymbol{H}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{R}^N)$, respectively. Note that μ_j is discrete while Φ is Gaussian distributed.

Theorem 4: Suppose for any closed and bounded set $\mathbb{V}$ of $\mathbb{R}^N$, we have $\mathbb{V} \subseteq \mathbb{S}_j$ for all j sufficiently large. Then $\{\mu_j\}_{j=1}^{\infty}$ converges to Φ in *Lévy* metric.

Sketch of Proof: First, consider the special case when $\boldsymbol{H}$ is an $N \times N$ diagonal matrix and $\Delta = \boldsymbol{I}_N$. For clarity, we use $\boldsymbol{D}$ to denote $\boldsymbol{H}$ and χ_n to denote its diagonal entries $\boldsymbol{D}^{(n,n)}$ for $n \in \mathsf{N}$. Using the water-filling algorithm [18], [19], there exist $P_1, \ldots, P_N \geq 0$ and $\vartheta > 0$ such that 1) $\sum_{n \in \mathsf{N}} P_n = \gamma$, 2) $P_n = \max(0, \vartheta - 1/\chi_n^2)$, and 3) for all $n \in \mathsf{N}$, we have $\chi_n^2/(\chi_n^2 P_n + 1) \leq 1/\vartheta$ and $P_n\chi_n^2/(\chi_n^2 P_n + 1) = P_n/\vartheta$. Let Φ be the Gaussian input distribution of $\boldsymbol{X}$ such that the components of $\boldsymbol{X}$ are zero mean and independent with variances $P_1, \ldots, P_N$, respectively. Then Φ is capacity achieving for the channel $(\boldsymbol{I}_N, \boldsymbol{D}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{R}^N)$.

By Proposition 6, we show that $\lim_{j\to\infty} I(\mu_j) = I(\Phi)$, or equivalently, $\lim_{j\to\infty} H_{\boldsymbol{Y}}(\mu_j) = H_{\boldsymbol{Y}}(\Phi)$. Let Γ be the set of all input probability measures μ of $\boldsymbol{X}$ such that $\boldsymbol{E}_{\mu}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}] \leq \gamma$. In Lemma 10, we prove that Γ is sequentially compact. Hence, $\{\mu_j\}_{j=1}^{\infty}$ converges to Φ if all convergent subsequences of $\{\mu_j\}_{j=1}^{\infty}$ converge to Φ. Assume without loss of generality that $\{\mu_j\}_{j=1}^{\infty}$ converges to an input probability measure μ_0 in Γ. As $\lim_{j\to\infty} H_{\boldsymbol{Y}}(\mu_j) = H_{\boldsymbol{Y}}(\Phi)$, by Lemma 9, we have

$$\lim_{j\to\infty} \int |P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_j) - P_{\boldsymbol{Y}}(\boldsymbol{y}; \Phi)| d\boldsymbol{y} = 0.$$

On the other hand, as $\{\mu_j\}_{j=1}^{\infty}$ converges to μ_0

$$\lim_{j\to\infty} P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_j) = P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_0), \qquad \text{for all } \boldsymbol{y} \in \mathbb{R}^M.$$

Therefore, for any $L > 0$

$$\begin{aligned}\int_{\|\boldsymbol{y}\| \leq L} &|P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_0) - P_{\boldsymbol{Y}}(\boldsymbol{y}; \Phi)| d\boldsymbol{y} \\ &\leq \underline{\lim}_{j\to\infty} \left(\begin{array}{l} \int |P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_j) - P_{\boldsymbol{Y}}(\boldsymbol{y}; \Phi)| d\boldsymbol{y} \\ + \int_{\|\boldsymbol{y}\| \leq L} |P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_j) - P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_0)| d\boldsymbol{y} \end{array} \right) \\ &= 0.\end{aligned}$$

As L is arbitrary, by the continuity of $P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_0)$ and $P_{\boldsymbol{Y}}(\boldsymbol{y}; \Phi)$, we have $P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_0) = P_{\boldsymbol{Y}}(\boldsymbol{y}; \Phi)$ for all $\boldsymbol{y} \in \mathbb{R}^N$. Hence, $\boldsymbol{E}_{\mu_0}[Y_n^2] = \boldsymbol{E}_{\Phi}[Y_n^2]$ for all $n \in \mathsf{N}$. Let L be the subset $\{n \in \mathsf{N} : \chi_n \neq 0\}$. As $\boldsymbol{E}_{\mu_0}[Y_n^2] = \chi_n^2 \boldsymbol{E}_{\mu_0}[X_n^2] + 1$ and $\boldsymbol{E}_{\Phi}[Y_n^2] = \chi_n^2 \boldsymbol{E}_{\Phi}[X_n^2] + 1$, we have $\boldsymbol{E}_{\Phi}[X_n^2] = \boldsymbol{E}_{\mu_0}[X_n^2]$ for all $n \in \mathsf{L}$. Consequently

$$\gamma = \sum_{n \in \mathsf{L}} \boldsymbol{E}_{\Phi}[X_n^2] = \sum_{n \in \mathsf{L}} \boldsymbol{E}_{\mu_0}[X_n^2]$$

which implies that $\boldsymbol{E}_{\mu_0}[X_n^2] = 0$ for $n \notin \mathsf{L}$. In other words, with respect to μ_0 and Φ, X_n is deterministic and equal to zero for $n \notin \mathsf{L}$.

Let $P_{\boldsymbol{D}\boldsymbol{X}}(\mu_0)$ and $P_{\boldsymbol{D}\boldsymbol{X}}(\Phi)$ be the probability measures of $\boldsymbol{D}\boldsymbol{X}$ when the input probability measures of $\boldsymbol{X}$ are μ_0 and Φ, respectively. Since $P_{\boldsymbol{Y}}(\boldsymbol{y}; \mu_0) = P_{\boldsymbol{Y}}(\boldsymbol{y}; \Phi)$, by Lemma 4, $P_{\boldsymbol{D}\boldsymbol{X}}(\mu_0) = P_{\boldsymbol{D}\boldsymbol{X}}(\Phi)$. As X_n is deterministic with respect to μ_0 and Φ for those n such that $\chi_n = 0$, $P_{\boldsymbol{D}\boldsymbol{X}}(\mu_0) = P_{\boldsymbol{D}\boldsymbol{X}}(\Phi)$ implies that $\mu_0 = \Phi$. Hence, Theorem 4 holds for this simple case.

Now, consider a general channel $(\Delta, \boldsymbol{H}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{S})$. Let M' be the rank of $\boldsymbol{H}$. Using singular-value decomposition, we can construct an $N \times N$ orthonormal matrix $\boldsymbol{U}$, an $M' \times N$ diagonal matrix $\boldsymbol{D}^*$, an $M' \times M$ matrix $\boldsymbol{V}$, and an $M \times M$ matrix $\boldsymbol{B}$ such that 1) $I(\boldsymbol{X}; \boldsymbol{H}\boldsymbol{X} + \boldsymbol{Z}) = I(\boldsymbol{X}; \boldsymbol{V}\boldsymbol{B}\boldsymbol{H}\boldsymbol{X} + \boldsymbol{V}\boldsymbol{B}\boldsymbol{Z})$, 2) the rank of $\boldsymbol{V}\boldsymbol{B}\boldsymbol{H}$ is M', 3) $\boldsymbol{V}\boldsymbol{B}\boldsymbol{H} = \boldsymbol{D}^*\boldsymbol{U}$, and 4) $\boldsymbol{E}[\boldsymbol{V}\boldsymbol{B}\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{V}^{\mathsf{T}}] = \boldsymbol{I}_{M'}$, Then the channel is equivalent to $\boldsymbol{Y}^* = \boldsymbol{D}^*\boldsymbol{X}^* + \boldsymbol{Z}^*$ where $\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{X}$, $\boldsymbol{Z}^* = \boldsymbol{V}\boldsymbol{B}\boldsymbol{Z}$, and is denoted by the tuple $(\boldsymbol{I}_{M'}, \boldsymbol{D}^*\boldsymbol{x}^*, \{\|\boldsymbol{x}^*\|^2 - \gamma\}, \mathbb{S}^*)$ where $\mathbb{S}^* = \{\boldsymbol{U}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{S}\}$.

By adding $N - M'$ "dummy" output variables, the channel $(\boldsymbol{I}_{M'}, \boldsymbol{D}^*\boldsymbol{x}^*, \{\|\boldsymbol{x}^*\|^2 - \gamma\}, \mathbb{S}^*)$ is equivalent to $(\boldsymbol{I}_N, \widehat{\boldsymbol{D}}\boldsymbol{x}^*, \{\|\boldsymbol{x}^*\|^2 - \gamma\}, \mathbb{S}^*)$, where $\widehat{\boldsymbol{D}}$ is obtained by appending $N - M'$ rows of zeros in $\boldsymbol{D}^*$. It is obvious that μ_j and Φ are still capacity-achieving measures for the channels $(\boldsymbol{I}_N, \widehat{\boldsymbol{D}}\boldsymbol{x}^*, \{\|\boldsymbol{x}^*\|^2 - \gamma\}, \mathbb{S}_j^*)$ and $(\boldsymbol{I}_N, \widehat{\boldsymbol{D}}\boldsymbol{x}^*, \{\|\boldsymbol{x}^*\|^2 - \gamma\}, \mathbb{R}^N)$ where $\mathbb{S}_j^* = \{\boldsymbol{U}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{S}_j\}$. Since $\widehat{\boldsymbol{D}}$ is a diagonal square matrix, using the same argument as before, we prove that $\{\mu_j\}_{j=1}^{\infty}$ converges to Φ. □

IV. Examples

A. Optical Channel—PAM

Consider an optical channel using intensity modulation and direct detection, in which an electrical signal is directly con-

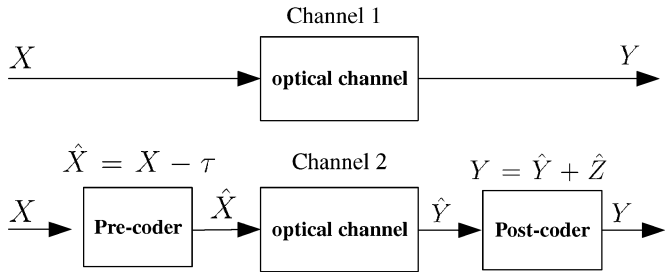

Fig. 1. Two optical channels.

verted to an optical intensity signal (e.g., by a laser diode). A discrete-time model for this channel is $Y = X + Z$ where Z is a zero-mean additive Gaussian noise, whose power may depend on the channel input. Specifically, we assume that $\Delta_x = \sigma_0^2 + \sigma_1^2 x$ for $\sigma_0^2 > 0$ and $\sigma_1^2 \geq 0$. This system is subject to three channel input constraints: a) the average power constraint, b) the peak power constraint, and c) the nonnegativity constraint (due to the nonnegativity of intensity). Such a channel can be characterized by the tuple $(\sigma_0^2 + \sigma_1^2 x, x, \{x - \gamma\}, [0, s])$.

By Theorem 1, there exists a probability measure μ_o in Λ which solves the channel capacity problem. In fact, by Theorem 2, there exists $\lambda \geq 0$ such that for all $x \in [0, s]$

$$Q(x; \mu_o) - \frac{1}{2}\log[2\pi e(\sigma_0^2 + \sigma_1^2 x)] - I(\mu_o) - \lambda(x - \gamma) \leq 0. \tag{16}$$

Furthermore, the inequality (16) is satisfied with equality for all $x \in \mathbb{E}_{\mu_o}$.

Corollary 2: Let μ_o be the capacity-achieving measure for the channel $(\sigma_0^2 + \sigma_1^2 x, x, \{x - \gamma\}, [0, s])$. Then μ_o is a discrete distribution with a finite number of probability mass points.

Proof: Suppose $\sigma_1^2 > 0$. Since $\mathbb{S}$ is connected, it is contained in one of the maximal connected open subsets of $\mathbb{A}(\Delta_w)$. Denote this connected open subset by $\mathbb{W}$. As $\Delta_w = \sigma_0^2 + \sigma_1^2 w$ and $\boldsymbol{b}_w = w$, the interval $(-\sigma_0^2/\sigma_1^2, +\infty)$ is a subset of $\mathbb{W}$. Let $\{w^{(i)}\}_{i=1}^{\infty}$ be a convergent sequence in the interval $(-\sigma_0^2/\sigma_1^2, +\infty)$ with limit $-\sigma_0^2/\sigma_1^2$. Then $\lim_{i\to\infty} \det \Delta_{w^{(i)}} = 0$ and for all positive integers i, $\boldsymbol{b}_{w^{(i)}}$ and $\Delta_{w^{(i)}}$ are real. Clearly, the requirements in Case B of Theorem 3 are satisfied, and thus, μ_o is discrete.

In the case $\sigma_1^2 = 0$, the channel is constant and is subject to a linear average cost constraint. Hence, the requirements in Case C of Theorem 3 are satisfied, and the capacity-achieving probability measure μ_o is thus discrete.

In both cases, $\mathbb{E}_{\mu_o}$, the set of points of increase of μ_o, is a sparse set. Since $\mathbb{E}_{\mu_o}$ is a bounded subset of $\mathbb{R}$, it is a finite set, i.e., μ_o has a finite number of probability mass points. □

Proposition 1: Suppose μ_o is the capacity-achieving measure for the channel $(\sigma_0^2 + \sigma_1^2 x, x, \{x - \gamma\}, [0, s])$. Then $x = 0$ is a point of increase of μ_o.

Proof: Suppose, to the contrary, that $x = 0$ is not a point of increase of μ_o. Let τ be the minimal element in $\mathbb{E}_{\mu_o}$, and hence $\tau > 0$. Consider the two channels depicted in Fig. 1. Channel 1 is the original optical channel, and channel 2 is obtained from channel 1 by appending a "pre-coder" and a "post-coder" before and after the inner optical channel. Specifically, $\hat{X} = X - \tau$ and $Y = \hat{Y} + \hat{Z}$ where $\hat{Z}$ is an independent additive Gaussian noise $\hat{Z}$ with mean τ and covariance $\tau\sigma_1^2$.

For any $x \geq \tau$, the conditional probability density functions of Y given $X = x$ is the same in both channels. Hence, if the input probability measure of the two channels is μ_o, then the joint probability measures of X and Y in the two channels are still the same, and consequently, so is the mutual information between the channel input and output.

In the second channel, as X, $\hat{X}$, $\hat{Y}$, and Y form a Markov chain $X \longrightarrow \hat{X} \longrightarrow \hat{Y} \longrightarrow Y$, we have $I(\hat{X}; \hat{Y}) \geq I(X; Y)$ by data processing inequality. Let ν be the corresponding probability measure of $\hat{X}$ when the probability measure of X is μ_o. Clearly, ν satisfies the bounded-input constraint and the average power constraint. Thus, ν is also capacity-achieving for channel 1, and hence, $P_Y(y; \mu_o) = P_Y(y; \nu)$ by Theorem 1. As a result, for channel 2, given the input probability measure of X is μ_o, the probability density function for Y and $\hat{Y}$ are the same, which is not possible since $\boldsymbol{E}[Y] = \boldsymbol{E}[\hat{Y}] + \tau$. Hence, the proposition follows. □

Corollary 3: If μ_o is capacity achieving for the channel $(\sigma_0^2 + \sigma_1^2 x, x, \{x - \gamma\}, [0, s])$, and it satisfies (16), then

$$\lambda = \left[I(\mu_o) - Q(0; \mu_o) + \frac{1}{2}\log 2\pi e \sigma_0^2 \right] / \gamma.$$

Proof: Since 0 is a point of increase of μ_o, the inequality (16) is tight at 0. Therefore, the Lagrange multiplier λ can be obtained by putting $x = 0$ in (16). □

For each $b \in \{2, 3, \ldots\}$, let $\tau^{(b)}$ be an input probability measure in Λ that maximizes $I(\mu)$ and has b or fewer points of increase. Using an approach similar to that of [1], the capacity-achieving measure of this optical channel can be found via the following search algorithm.

Search Algorithm for Capacity-Achieving Measures

Step 1: Set $b = 2$.

Step 2: Solve for $\tau^{(b)}$.

Step 3: Let $\lambda^{(b)} = [I(\tau^{(b)}) - Q(0; \tau^{(b)}) + \log 2\pi e \sigma_0^2]/\gamma$. If $\lambda^{(b)} < 0$, increase b by 1 and go back to step 2.

Step 4: Verify whether the inequality

$$Q(x; \tau^{(b)}) - I(\tau^{(b)}) - \frac{1}{2}\log[2\pi e(\sigma_0^2 + \sigma_1^2 x)] - \lambda^{(b)}(x - \gamma) \leq 0$$

holds for all $x \in [0, s]$. If so, then $\tau^{(b)}$ is capacity achieving. If otherwise, increase b by 1 and go back to step 2.

Numerical Results

The parameter σ_1^2 determines the degree of signal dependency of the optical channel. It is interesting to know how the parameter σ_1^2 affects the capacity-achieving measure. In the above optical channel, we fix the following parameters: $\sigma_0^2 = 0.1$, $\gamma = 2$, and $s = 6$. Let $\mu[\sigma_1^2]$ be the capacity-achieving measure for the channel $(\sigma_0^2 + \sigma_1^2 x, x, \{x - \gamma\}, [0, s])$. With respect to different specified values of σ_1^2, via the above search algorithm, $\mu[\sigma_1^2]$ was found and its set of points of increases are plotted in Fig. 2. Specifically, the horizontal axis denotes the values of σ_1^2 and the vertical axis denotes the input signal. If a dot is indicated in the

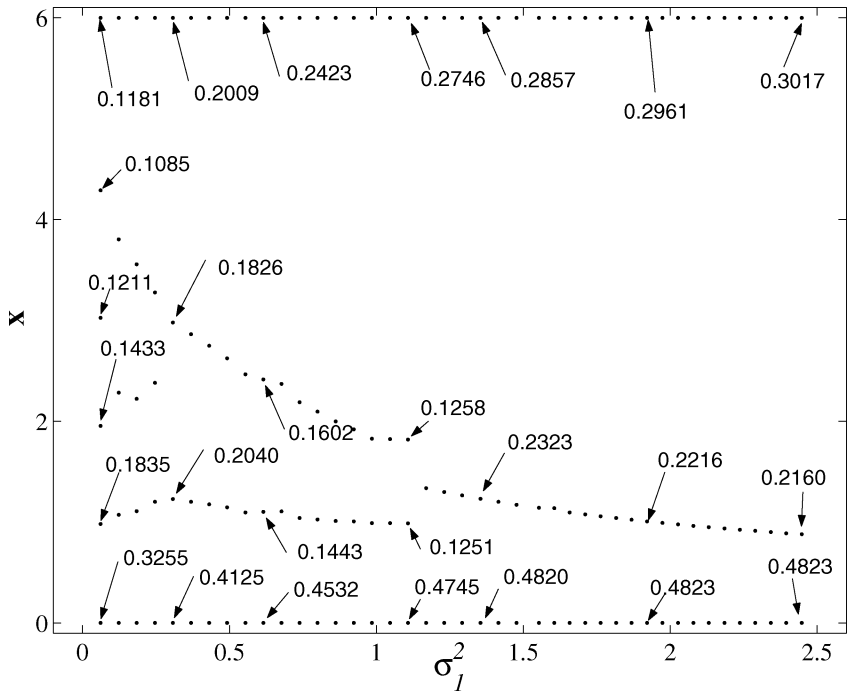

Fig. 2. Capacity-achieving measures.

position (σ_1^2, x^*), then x^* is a point of increase of $\mu[\sigma_1^2]$. Moreover, the probabilities at the points of increase of $\mu[\sigma_1^2]$ are also shown in the figure for selected values of σ_1^2.

From the figure, we observe the following.

1. 0 and s are always points of increase of $\mu[\sigma_1^2]$. It is proved in Proposition 1 that 0 is always a point of increase of $\mu[\sigma_1^2]$. However, it is not known whether s is also a point of increase of $\mu[\sigma_1^2]$ in general.
2. The distance between two neighboring points of increases of $\mu[\sigma_1^2]$ varies significantly. In particular, the separation is smaller if the interval is "closer" to 0. This phenomenon is especially apparent when the channel is highly signal dependent (i.e., when σ_1^2 is large). We believe that this phenomenon is due to the fact that when σ_1^2 is large, input signals of larger power require greater separation than those of smaller power.
3. As σ_1^2 increases, the size of $\mathbb{E}_{\mu[\sigma_1^2]}$ decreases. This happens because when σ_1^2 increases, the average noise power also increases; hence, a signaling scheme with a smaller constellation is more favorable.
4. The probabilities of "$x = 0$" and "$x = s$" increase, as σ_1^2 increases. For a fixed channel input, when σ_1^2 increases, the induced noise power also increases. Hence, for large σ_1^2, it is more favorable to use inputs of smaller power. This explains why the probability of "$x = 0$" increases. An unexpected observation from our numerical results is that the probability of "$x = s$" also increases. We believe that this is because the increase in probabilities of "$x = s$" and "$x = 0$" also increases the "average distances" between channel codewords. According to our numerical results, this advantage seems to outweigh the disadvantage of the increase in average noise power.

B. Fading Channel

Consider a general multiple-antenna system operated in a fading environment. The channel input and output are complex-valued $\widetilde{N}$-tuple $\widetilde{\boldsymbol{X}}$ and $\widetilde{M}$-tuple $\widetilde{\boldsymbol{Y}}$, respectively, such that $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ are related as $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{X}} + \widetilde{\boldsymbol{Z}}$. Here, we do not make any assumption on the complex channel gain matrix $\widetilde{\boldsymbol{H}}$ and the additive noise $\widetilde{\boldsymbol{Z}}$ except that 1) they are complex Gaussian distributed and 2) $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{Z}}$ are independent to each other. Such a channel can also be formulated as a quadratic CG channel $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{Z}$, where

$$\boldsymbol{Y} = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{Y}}) \\ \mathrm{im}(\widetilde{\boldsymbol{Y}}) \end{bmatrix}, \quad \boldsymbol{H} = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{H}}) & -\mathrm{im}(\widetilde{\boldsymbol{H}}) \\ \mathrm{im}(\widetilde{\boldsymbol{H}}) & \mathrm{re}(\widetilde{\boldsymbol{H}}) \end{bmatrix}$$
$$\boldsymbol{X} = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{X}}) \\ \mathrm{im}(\widetilde{\boldsymbol{X}}) \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{Z}}) \\ \mathrm{im}(\widetilde{\boldsymbol{Z}}) \end{bmatrix}. \tag{17}$$

Example: Rayleigh-Fading Channels: Consider the Rayleigh-fading channel, in which all the entries in $\widetilde{\boldsymbol{H}}$ and $\widetilde{\boldsymbol{Z}}$ are independent, zero-mean, complex Gaussian distributed with variance χ^2 and σ^2, respectively. Hence, for any input $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{b}_{\boldsymbol{x}} = \boldsymbol{0}$ and

$$\Delta_{\boldsymbol{x}} = \frac{1}{2}\left(\sum_{n\in\mathsf{N}} \chi^2 x_n^2 + \sigma^2\right) \boldsymbol{I}_M.$$

Subject to peak and average total power constraints, the channel is denoted by the tuple

$$\left(\frac{1}{2}\left(\sum_{n\in\mathsf{N}} \chi^2 x_n^2 + \sigma^2\right)\boldsymbol{I}_M, \boldsymbol{0}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{B}_N(s^2)\right)$$

where s^2 and γ are the maximal peak and average total power respectively.

Corollary 4: Let μ_o be the capacity-achieving measure for the above Rayleigh-fading channel. Then μ_o is discrete. In particular, its set of points of increase is contained in a finite number

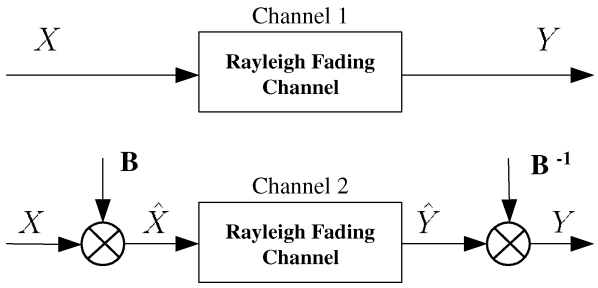

Fig. 3. Two equivalent Rayleigh-fading channels.

of "concentric shells" centered at $\mathbf{0}$. In other words, the probability distribution of $||\boldsymbol{X}||$ is discrete with a finite number of probability mass points.

Proof: Let $\mathbb{W}$ be the maximal connected open subset of $\mathbb{A}(\Delta_{\boldsymbol{w}})$ containing $\mathbb{B}_N(s^2)$. Then, for any $0 \le \epsilon < \sigma/\chi$, the vector $(\boldsymbol{j}\epsilon, 0, \ldots, 0) \in \mathbb{W}$. Let $\{\boldsymbol{w}^{(i)} \triangleq (\boldsymbol{j}\epsilon_i, 0, \ldots, 0)\}_{i=1}^{\infty}$ be a sequence in $\mathbb{W}$ such that $\lim_{i\to\infty} \epsilon_i = \sigma/\chi$ and $0 \le \epsilon_i < \sigma/\chi$ for all positive integers i. Obviously, $\lim_{i\to\infty} \det \Delta_{\boldsymbol{w}^{(i)}} = 0$ and for all positive integers i, $\boldsymbol{b}_{\boldsymbol{w}^{(i)}}$ and $\Delta_{\boldsymbol{w}^{(i)}}$ are real. The requirements in Case B of Theorem 3 are thus satisfied, and hence, μ_o is discrete. It remains to prove that $\mathbb{E}_{\mu_o}$ is a subset of a finite number of concentric shells centered at $\mathbf{0}$.

Consider the two channels depicted in Fig. 3. Channel 1 is the Rayleigh-fading channel, and channel 2 is obtained from channel 1 by multiplying a real orthonormal matrix $\boldsymbol{B}$ and its inverse $\boldsymbol{B}^{-1}$ before and after the Rayleigh-fading channel. It is easy to see that for any $\boldsymbol{x} \in \mathbb{R}^N$, the conditional probability density functions of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ is the same for both channels. Hence, if the input probability measures of the two channels are equal to μ_o, then the mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ in both channels is the same.

In the second channel, as $\boldsymbol{X}$, $\hat{\boldsymbol{X}}$, $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ form a Markov chain $\boldsymbol{X} \longrightarrow \hat{\boldsymbol{X}} \longrightarrow \hat{\boldsymbol{Y}} \longrightarrow \boldsymbol{Y}$, we have $I(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}}) \ge I(\boldsymbol{X};\boldsymbol{Y})$. Let ν be the probability measure of $\hat{\boldsymbol{X}}$. It is obvious that ν satisfies all channel input constraints. Thus, ν is also capacity achieving, and $I(\mu_o) = I(\nu)$.

By Theorem 2, there exists $\lambda \ge 0$ such that for all $\boldsymbol{x} \in \mathbb{B}_N(s^2)$

$$Q(\boldsymbol{x};\mu_o) - \frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{x}}] - I(\mu_o) - \lambda(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma) \le 0. \tag{18}$$

Let $\mathbb{E}$ be the subset of $\mathbb{B}_N(s^2)$ such that (18) is tight. If $\boldsymbol{x} \in \mathbb{E}_{\mu_o}$, then $\boldsymbol{B}\boldsymbol{x} \in \mathbb{E}_{\nu}$. Hence, $\boldsymbol{x}$ and $\boldsymbol{B}\boldsymbol{x}$ are both in $\mathbb{E}$. Since $\boldsymbol{B}$ is an arbitrary orthonormal matrix, $\mathbb{E}$ contains the set

$$\{\boldsymbol{x} \in \mathbb{R}^N : \exists \boldsymbol{w} \in \mathbb{E}_{\mu_o} \text{ such that } ||\boldsymbol{x}||^2 = ||\boldsymbol{w}||^2\}$$

which is a collection of concentric shells centered at $\mathbf{0}$. By Lemma 1, the sparseness of $\mathbb{E}$ implies that the number of concentric shells is finite and the corollary then follows. □

As a remark, the capacity-achieving measure for the Rayleigh-fading channel has been studied separately in [4], in which it is proved that the capacity-achieving measure is discrete even when there is no peak power constraint.

Suppose the receiver can estimate the channel correctly. In other words, suppose the channel matrix $\boldsymbol{H}$ is known to the receiver. Let $\hat{\boldsymbol{H}}$ be an $MN \times 1$ column vector such that it contains all the entries of $\boldsymbol{H}$. Such a fading channel with receiver-side channel information can still be formulated as a CG channel, where the output of the channel is $[\hat{\boldsymbol{H}}^{\mathsf{T}}, \boldsymbol{Y}^{\mathsf{T}}]^{\mathsf{T}}$. It is easy to prove that for all $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{b}_{\boldsymbol{x}}$ and $\Delta_{\boldsymbol{x}}$ are real,

$$\mathrm{tr}(\Delta_{\boldsymbol{x}}) = MN\chi^2/2 + M\sigma^2/2 + M||\boldsymbol{x}||^2\chi^2/2,$$

and

$$\log\det\Delta_{\boldsymbol{x}} = MN\log(\chi^2/2) + M\log(\sigma^2/2), \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^N.$$

Corollary 5: For the above Rayleigh-fading channel with channel side information available at the receiver, subject only to the peak total power constraint,[3] there exists a capacity-achieving measure μ_o which is DAUP.

Proof: Let μ^* be a capacity-achieving measure and $\mathbb{W}$ be the maximal connected open subset of $\mathbb{A}(\Delta_{\boldsymbol{w}})$ containing $\mathbb{R}^N$. As $\mathrm{tr}(\Delta_{\boldsymbol{x}})$ is unbounded over $\mathbb{R}^N$, the requirements of Case A in Theorem 3 are satisfied. Hence, μ^* is discrete.

For any orthonormal matrix $\boldsymbol{B}$, let $\hat{\boldsymbol{X}} = \boldsymbol{B}\boldsymbol{X}$ and ν be its probability measure. Applying a similar technique when we proved Corollary 4, we can show that ν is also capacity-achieving. Since $I(\mu)$ is concave, any linear combination of μ^* and ν is still capacity achieving. By "averaging" over all possible orthonormal matrices $\boldsymbol{B}$, we can construct from μ^* a capacity-achieving measure μ_o which is uniform in phase. The phase uniformity of μ_o implies that $\mathbb{E}_{\mu_o}$ consists of a number of concentric shells centered at $\mathbf{0}$. Hence, the sparseness of $\mathbb{E}_{\mu_o}$ implies that the number of concentric shells in $\mathbb{E}_{\mu_o}$ is finite. In other words, μ_o is DAUP. □

Example: MIMO Rayleigh Block-Fading Channels: Consider a MIMO Rayleigh block-fading channel [15], whose channel gain matrix remains constant for T symbol periods, after which it changes to a new set of values independently, and again, maintain for another T symbol periods, and so on. Let $\widetilde{N}$ and $\widetilde{M}$ be the number of transmitting and receiving antennas, respectively. Within a block of T channel input vectors, for $l = 1, \ldots, T$, let

$$\widetilde{\boldsymbol{X}}^{(l)} = \left[\widetilde{X}_1^{(l)}, \ldots, \widetilde{X}_{\widetilde{N}}^{(l)}\right]^{\mathsf{T}}$$

be the lth channel input vector and

$$\widetilde{\boldsymbol{Y}}^{(l)} = \left[\widetilde{Y}_1^{(l)}, \ldots, \widetilde{Y}_{\widetilde{M}}^{(l)}\right]^{\mathsf{T}}$$

be the corresponding channel output vector. Then $\widetilde{\boldsymbol{X}}^{(l)}$ and $\widetilde{\boldsymbol{Y}}^{(l)}$ are related by the equation $\widetilde{\boldsymbol{Y}}^{(l)} = \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{X}}^{(l)} + \widetilde{\boldsymbol{Z}}^{(l)}$ where $\widetilde{\boldsymbol{H}}$ is the complex channel gain matrix for the Rayleigh channel as defined in Section IV-B and $\widetilde{\boldsymbol{Z}}^{(l)}$ is the corresponding additive complex Gaussian noise. For simplicity, let

$$\boldsymbol{X}^{(l)} = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{X}}^{(l)}) \\ \mathrm{im}(\widetilde{\boldsymbol{X}}^{(l)}) \end{bmatrix}, \quad \boldsymbol{Y}^{(l)} = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{Y}}^{(l)}) \\ \mathrm{im}(\widetilde{\boldsymbol{Y}}^{(l)}) \end{bmatrix}$$

[3]The peak total power is the sum of the powers of all transmitting antennae

$$\boldsymbol{Z}^{(l)} = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{Z}}^{(l)}) \\ \mathrm{im}(\widetilde{\boldsymbol{Z}}^{(l)}) \end{bmatrix}, \ \boldsymbol{H}^* = \begin{bmatrix} \mathrm{re}(\widetilde{\boldsymbol{H}}) & -\mathrm{im}(\widetilde{\boldsymbol{H}}) \\ \mathrm{im}(\widetilde{\boldsymbol{H}}) & \mathrm{re}(\widetilde{\boldsymbol{H}}) \end{bmatrix}$$
$$\boldsymbol{X} = \left[(\boldsymbol{X}^{(1)})^{\mathsf{T}}, \ldots, (\boldsymbol{X}^{(T)})^{\mathsf{T}}\right]^{\mathsf{T}}$$
$$\boldsymbol{Y} = \left[(\boldsymbol{Y}^{(1)})^{\mathsf{T}}, \ldots, (\boldsymbol{Y}^{(T)})^{\mathsf{T}}\right]^{\mathsf{T}}$$
$$\boldsymbol{Z} = \left[(\boldsymbol{Z}^{(1)})^{\mathsf{T}}, \ldots, (\boldsymbol{Z}^{(T)})^{\mathsf{T}}\right]^{\mathsf{T}}. \tag{19}$$

It is easy to show that the MIMO Rayleigh block-fading channel can be rewritten as a quadratic CG channel $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{Z}$ where $\boldsymbol{H}$ is a $2\widetilde{M}T \times 2\widetilde{N}T$ "block-diagonal matrix" whose diagonal blocks are all equal to $\boldsymbol{H}^*$. Since the channel is quadratic, the well-behaved region $\mathbb{A}(\Delta_{\boldsymbol{w}})$ contains $\mathbb{R}^{2T\widetilde{N}}$. As $\boldsymbol{H}$ is zero mean, $\boldsymbol{b}_{\boldsymbol{x}} = \boldsymbol{0}$ for all $\boldsymbol{x} \in \mathbb{R}^{2\widetilde{N}T}$.

Corollary 6: Let μ_o be the capacity-achieving measure for the above Rayleigh block-fading channel subject to peak and average power constraint. Then μ_o is discrete.

Proof: Let $\mathbb{W}$ be the maximal connected open subset of $\mathbb{A}(\Delta_{\boldsymbol{w}})$ that contains $\mathbb{R}^{2\widetilde{N}T}$. Let $\boldsymbol{w}^* = (\boldsymbol{j}\epsilon, 0, \ldots, 0)$ be a vector in $\mathbb{C}^{2\widetilde{N}T}$ where $0 \le \epsilon < \sigma/\chi$. By direct substitution, it can be verified easily that $\boldsymbol{w}^* \in \mathbb{W}$. Furthermore, $\Delta_{\boldsymbol{w}^*}$ are real and

$$\det \Delta_{\boldsymbol{w}^*} = \left(\frac{\sigma^2 - \chi^2\epsilon^2}{2}\right)^{2\widetilde{M}} \left(\frac{\sigma^2}{2}\right)^{2(T-1)\widetilde{M}}. \tag{20}$$

Let $\{\boldsymbol{w}^{(i)} \triangleq (\boldsymbol{j}\epsilon_i, 0, \ldots, 0)\}_{i=1}^{\infty}$ be a sequence in $\mathbb{W}$ such that $\lim_{i\to\infty} \epsilon_i = \sigma/\chi$ and $0 \le \epsilon_i < \sigma/\chi$ for all positive integers i. Obviously, $\lim_{i\to\infty} \det \Delta_{\boldsymbol{w}^{(i)}} = 0$ and for all positive integers i, $\boldsymbol{b}_{\boldsymbol{w}^{(i)}}$ and $\Delta_{\boldsymbol{w}^{(i)}}$ are real. The requirements in Case B of Theorem 3 are thus satisfied, and hence, μ_o is discrete. □

As a remark, the capacity-achieving measure for the single-antenna Rayleigh block-fading channel has been studied separately in [8], in which it is proved that the capacity-achieving measure is discrete even when there is no peak power constraint.

Example: Rician Fading Channels: The Rician fading channel is a generalization of the Rayleigh-fading channel in the sense that $\boldsymbol{E}[\boldsymbol{H}]$ is not necessarily a zero matrix as in the case of Rayleigh-fading channel. Specifically, given the channel input is $\boldsymbol{x}$,

$$\Delta_{\boldsymbol{x}} = \frac{1}{2}\left(\sum_{n\in\mathsf{N}} \chi^2 x_n^2 + \sigma^2\right)$$

and $\boldsymbol{b}_{\boldsymbol{x}} = \boldsymbol{E}[\boldsymbol{H}]\boldsymbol{x}$ which is nonzero in general.

Corollary 7: The capacity-achieving probability measure for the Rician fading channel

$$\left(\frac{1}{2}\left(\sum_{n\in\mathsf{N}} \chi^2 x_n^2 + \sigma^2\right) \boldsymbol{I}_M, \boldsymbol{E}[\boldsymbol{H}]\boldsymbol{x}, \{\}, \mathbb{B}_N(s^2)\right)$$

is discrete.

Proof: Since a Rician fading channel is a quadratic CG channel, by Corollary 1, the result follows. □

As a remark, many fading channels can be formulated as quadratic CG channels. Therefore, according to Corollary 1, their capacity-achieving measures are discrete if there is no average cost constraint.

C. Example: Interference Channels

Consider a communication system consisting of N pairs of transmitters and receivers. Each pair is connected by a point-to-point subchannel, and the subchannels interfere with each other. Specifically, the channel input $\boldsymbol{X}$ and output $\boldsymbol{Y}$ are related as $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{Z}$ such that 1) the entries of $\boldsymbol{H}$ and $\boldsymbol{Z}$ are Gaussian distributed independently with variances χ^2 and σ^2, respectively, 2) $\boldsymbol{Z}$ is zero mean and $\boldsymbol{E}[\boldsymbol{H}]$ is the identity matrix. Therefore, for any input $\boldsymbol{x}$, $\boldsymbol{b}_{\boldsymbol{x}} = \boldsymbol{x}$, and

$$\Delta_{\boldsymbol{x}} = \left(\sum_{n\in\mathsf{N}} \chi^2 x_n^2 + \sigma^2\right) \boldsymbol{I}_N.$$

Corollary 8: There exists a DAUP capacity-achieving measure μ_o for the interference channel

$$\left(\left(\sum_{n\in\mathsf{N}} \chi^2 x_n^2 + \sigma^2\right) \boldsymbol{I}_N, \boldsymbol{x}, \{\}, \mathbb{B}_N(s^2)\right).$$

Proof: Since the channel is quadratic and there is no average cost constraint, the capacity-achieving measure is discrete. The proof of that μ_o is DAUP is the same as the one given in Corollary 5. □

D. Example: Parallel Gaussian Channels

Consider a system of N independent and identically scalar additive Gaussian channels in parallel subject to an average total power constraint $\boldsymbol{E}_\mu[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} - \gamma] \le 0$ and a bounded-input constraint. In other words, the channel is characterized by the tuple $(\sigma^2\boldsymbol{I}_N, \boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{S})$. By Theorems 1 and 3, the capacity-achieving measure is unique and discrete. We will consider two cases of bounded-input constraints: cubic and spherical constraints.

Case A: Cubic Constraint: In the first case, the set of admissible channel input vectors is a hypercube, defined as $\mathbb{S} = \prod_{n\in\mathsf{N}} \mathbb{S}_n$ where $\mathbb{S}_n = \{x_n \in \mathbb{R} : x_n^2 \le s^2\}$. In other words, the peak power of each component channel is s^2. For any input probability measure μ of $\boldsymbol{X}$, we denote the corresponding marginal probability measure of X_n by $\mu^{(n)}$.

Let μ_o be the capacity-achieving measure for the channel $(\sigma^2\boldsymbol{I}_N, \boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \prod_{n\in\mathsf{N}} \mathbb{S}_n)$ and ν be the product measure of $\{\mu_o^{(1)}, \ldots, \mu_o^{(N)}\}$. It can be proved easily that $I(\nu) \ge I(\mu_o)$ and $\nu \in \Lambda$. By the uniqueness of μ_o, we have $\mu_o = \nu$ and $\mu^{(1)} = \mu^{(2)} = \cdots = \mu^{(N)}$. By direct substitution, we can verify immediately that $I(\nu) = NC$ where C is the channel capacity of the channel $(\sigma^2, x_n, \{x_n^2 - \gamma/N\}, [-s, s])$ and $\mu_o^{(n)}$ is the corresponding capacity-achieving measure. Again, by Theorem 3, $\mu_o^{(n)}$ is discrete and hence has a finite set of points of increase. As $\mathbb{E}_{\mu_o} = \prod_{n\in\mathsf{N}} \mathbb{E}_{\mu_o^{(n)}}$, $\mathbb{E}_{\mu_o}$ is also finite.

Case B: Spherical Constraint: In the second case, the set of admissible channel input vectors is $\mathbb{B}_N(s^2)$. Let μ_o be the capacity-achieving measure for the channel $(\sigma^2\boldsymbol{I}_N, \boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{B}_N(s^2))$. For any orthonormal matrix $\boldsymbol{B}$, let $\hat{\boldsymbol{X}} = \boldsymbol{B}\boldsymbol{X}$ and ν be its probability measure. Again, using the same technique as in proving Corollary 4, we can show that ν is also capacity achieving. The uniqueness of the capacity-achieving measures thus implies that $\mu_o = \nu$. Hence, μ_o is uniform in phase, and $\mathbb{E}_{\mu_o}$ consists of a number of concentric shells centered at $\boldsymbol{0}$. The sparseness of $\mathbb{E}_{\mu_o}$ thus implies that the number of concentric

shells in $\mathbb{E}_{\mu_o}$ is finite. In other words, μ_o is DAUP. The special case when $N = 2$ was proved by Shamai and Bar-David in [3].

V. Conclusion

In this paper, we have shown that, in many instances, the capacity-achieving probability measure for a conditionally Gaussian channel with bounded-input constraints is discrete. The criteria for discreteness given in Theorem 3 are not exhaustive, and hence, there may be many more examples of CG channels with a discrete capacity-achieving measure. In fact, according to Theorem 2, for a conditionally Gaussian channel subject to bounded-input constraints, if its capacity-achieving probability measure is not discrete, then the inequality in (6) must be tight for all admissible channel inputs $\boldsymbol{x}$. We believe that such a requirement is rather stringent in general. In other words, it is more natural to expect that most capacity-achieving measures are discrete.

There is much work yet to be done in the analysis of channels with input constraints. The tools we used allow us to verify whether a given input measure is capacity achieving or not. Once the capacity-achieving measure and the channel capacity are determined, they can serve as a benchmark to evaluate the efficiency of any practical scheme. Furthermore, the obtained capacity-achieving measure may lead to valuable insights in designing practical and efficient modulation schemes.

However, except in a few special cases, when we may guess the form (e.g., DAUP) of the capacity-achieving probability measure, these capacity-achieving probability measures are extremely difficult to obtain. More sophisticated techniques are yet to be discovered to overcome this problem.

Appendix I
Proof of Lemma 1

Suppose I is finite. Let $b(\boldsymbol{w}) = \prod_{i\in\mathsf{I}}(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} - p_i^2)$. Clearly, the function $b(\boldsymbol{w})$ is holomorphic over $\mathbb{C}^N$ and is zero in $\mathbb{F}$. Therefore, $\mathbb{F}$ is sparse. It remains to prove that if I is not finite, then $\mathbb{F}$ is not sparse.

Suppose I is not finite. The Bolzano–Weierstrass theorem implies that $\{p_i : i \in \mathsf{I}\}$ has a limiting point p. Let $b(\boldsymbol{w})$ be a holomorphic function defined on a connected open subset $\mathbb{U}$ of $\mathbb{C}^N$ containing the closure of $\mathbb{F}$ such that $b(\boldsymbol{w}) = 0$ for all $\boldsymbol{w} \in \mathbb{F}$.

First, consider the case when $p > 0$. Let $\boldsymbol{v}$ be any point in $\mathbb{R}^N$ such that $\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v} = p^2$. Then $\boldsymbol{v}$ is a limiting point of $\mathbb{F}$, and is contained in $\mathbb{U}$. Let $\mathbb{W}(\epsilon, \delta)$ be as in the equation at the bottom of the page. By picking $\epsilon, \delta > 0$ small enough, $\mathbb{W}(\epsilon, \delta)$ can be made to be a subset of $\mathbb{U}$.

Let $\boldsymbol{x} \in \mathbb{W}(\epsilon, \delta)$ be a vector such that $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} = p^2$. Let $f(v) = b(v\boldsymbol{x})$ be the single-variable complex-valued function, and $\widetilde{\mathbb{F}}$ be the set

$$\{\zeta \in (1-\delta, 1+\delta) : \exists p_i \text{ such that } \zeta = p_i/p \text{ for some } i \in \mathsf{I}\}.$$

It is easy to see that $f(v)$ is an analytic function on an open subset of $\mathbb{C}$ containing the interval $(1-\delta, 1+\delta)$, and $f(v)$ is zero on $\widetilde{\mathbb{F}}$. As p is a limiting point of $\{p_i : i \in \mathsf{I}\}$, 1 is a limiting point of $\widetilde{\mathbb{F}}$. By the identity theorem [20], $f(v) = 0$ for all $v \in (1-\delta, 1+\delta)$. In other words, $b(v\boldsymbol{x}) = 0$ for all $v \in (1-\delta, 1+\delta)$. As

$$\mathbb{W}(\epsilon, \delta) = \bigcup_{\boldsymbol{x}\in\mathbb{W}(\epsilon,\delta):\, \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x}=p^2} \{v\boldsymbol{x} :\, 1-\delta < v < 1+\delta\} \quad (21)$$

$b(\boldsymbol{w}) = 0$ for all $\boldsymbol{w} \in \mathbb{W}(\epsilon, \delta)$. Since $\mathbb{W}(\epsilon, \delta)$ is open in $\mathbb{R}^N$, again by the identity theorem, $b(\boldsymbol{w}) = 0$ for all $\boldsymbol{w} \in \mathbb{U}$. This implies that $\mathbb{F}$ is not sparse.

In the case when $p = 0$, then the origin $\mathbf{0}$ is a limiting point of $\mathbb{F}$. Let $\mathbb{W}(\delta) = \{\boldsymbol{u} \in \mathbb{R}^N : \|\boldsymbol{u}\| < \delta\}$. Picking $\delta > 0$ small enough, $\mathbb{W}(\delta)$ is a subset of $\mathbb{U}$. For any $\boldsymbol{x} \in \mathbb{R}^N$ such that $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} = 1$, let $f(v)$ be the single-variable complex-valued function $b(v\boldsymbol{x})$. Then $f(v)$ is analytic on an open subset of $\mathbb{C}$ containing the interval $(-\delta, \delta)$. As $\{p_i \in (-\delta, \delta) : i \in \mathsf{I}\}$ has a limiting point at 0 and $f(v)$ is zero on this set, $f(v) = 0$ for $v \in (-\delta, \delta)$, and consequently, $b(\boldsymbol{w}) = 0$ for all $\boldsymbol{w} \in \mathbb{W}(\delta)$. Again, by the identity theorem, $b(\boldsymbol{w}) = 0$ for all $\boldsymbol{w} \in \mathbb{U}$. The lemma is thus proved.

Appendix II
Supplementary Results

Let $\mathcal{B}_N$ be the vector space of all bounded and continuous real-valued functions defined on $\mathbb{R}^N$ and $\mathcal{B}_N^*$ be its dual space [21]. The set of input probability measures, denoted by Ω, can be identified as a subset of $\mathcal{B}_N^*$. In this paper, we assume that the topology on Ω is induced by the weak* topology defined on $\mathcal{B}_N^*$. The induced topology on Ω is metrizable, whose metric is called the *Lévy* metric [23]. This topology is complete and satisfies the following properties.

Lemma 2: Let $\{\mu_i\}_{i=0}^{\infty}$ be a sequence in Ω.

1. The sequence converges to μ_0 in the *Lévy* metric, denoted by $\mu_i \Longrightarrow \mu_0$, if and only if for every $f \in \mathcal{B}_N$
$$\lim_{i\to\infty} \int f(\boldsymbol{x}) d\mu_i = \int f(\boldsymbol{x}) d\mu_0.$$
In particular, if $\{\mu_i\}_{i=0}^{\infty}$ is a sequence in $\Lambda_{\mathbb{S}}$, then $\mu_i \Longrightarrow \mu_0$ if and only if for all-continuous function f defined on $\mathbb{S}$
$$\lim_{i\to\infty} \int f(\boldsymbol{x}) d\mu_i = \int f(\boldsymbol{x}) d\mu_0.$$
2. Suppose $\mu_i \Longrightarrow \mu_0$. For any closed subset $\mathbb{K}$ of $\mathbb{R}^N$,
$$\overline{\lim}_{i\to\infty} \mu_i(\mathbb{K}) \le \mu_0(\mathbb{K}).$$
Moreover, if $f(\boldsymbol{x})$ is continuous on $\mathbb{R}^N$ and is bounded below, then
$$\underline{\lim}_{i\to\infty} \int f d\mu_i \ge \int f d\mu_0.$$
3. A subset Υ of Ω is called **tight** if for any $\epsilon > 0$, there exists a closed and bounded subset $\mathbb{K}$ of $\mathbb{R}^N$ such that

$$\mathbb{W}(\epsilon, \delta) = \{\boldsymbol{u} \in \mathbb{R}^N :\, \|\boldsymbol{u}\|\|\boldsymbol{v}\|(1-\epsilon) < \boldsymbol{u}^{\mathsf{T}}\boldsymbol{v} \text{ and } (1-\delta)p < \|\boldsymbol{u}\| < (1+\delta)p\}.$$

$\mu(\mathbb{K}) \geq 1-\epsilon$ for all $\mu \in \Upsilon$. Suppose Υ is closed and tight. Then it is sequentially compact.

Proof: See [23, Sec. 3.1]. □

Proposition 2 (Convexity and Sequential Compactness): The subsets $\Lambda_{\mathbb{S}}$ and Λ of Ω are convex and sequentially compact.

Proof: The convexity of $\Lambda_{\mathbb{S}}$ and Λ are trivial. To prove that they are sequentially compact, it is sufficient to show that they are tight and closed. First, notice that $\mathbb{S}$ is closed and bounded. As $\mu(\mathbb{S})=1$ for all $\mu \in \Lambda_{\mathbb{S}}$, $\Lambda_{\mathbb{S}}$ and its subset Λ are tight.

Let $\{\mu_i\}_{i=1}^{\infty}$ be a convergent sequence in $\Lambda_{\mathbb{S}}$ with limit μ_0. Since $\mathbb{S}$ is closed, $\overline{\lim}_{i\to\infty}\mu_i(\mathbb{S}) \leq \mu_0(\mathbb{S})$, which further implies that $\mu_0(\mathbb{S})=1$, or, equivalently, $\mu_0 \in \Lambda_{\mathbb{S}}$. Therefore, $\Lambda_{\mathbb{S}}$ is closed, and thus sequentially compact.

Similarly, let $\{\mu_i\}_{i=1}^{\infty}$ be a convergent sequence in Λ with limit μ_0. The sequential compactness of $\Lambda_{\mathbb{S}}$ implies that $\mu_0 \in \Lambda_{\mathbb{S}}$. As $\mu_i \Longrightarrow \mu_0$ and $g_k(\boldsymbol{x})$ is continuous on $\mathbb{R}^N$ for all $k \in \mathsf{K}$, by Lemma 2, we have

$$0 \geq \lim_{i\to\infty}\int g_k(\boldsymbol{x})\,d\mu_i = \int g_k(\boldsymbol{x})\,d\mu_0.$$

Hence, $\mu_0 \in \Lambda$ and the sequential compactness of Λ then follows. □

Lemma 3: There exist positive constants α, β, β' such that for all $\boldsymbol{x} \in \mathbb{S}$ and $\boldsymbol{y} \in \mathbb{R}^M$

$$\exp(-\alpha-\beta'||\boldsymbol{y}||^2) \leq P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) \leq \exp(\alpha-\beta||\boldsymbol{y}||^2).$$

Hence, for all $\mu \in \Lambda_{\mathbb{S}}$

$$\exp(-\alpha-\beta'||\boldsymbol{y}||^2) \leq P_{\boldsymbol{Y}}(\boldsymbol{y};\mu) \leq \exp(\alpha-\beta||\boldsymbol{y}||^2)$$

and $|\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)| \leq \alpha+\beta'||\boldsymbol{y}||^2$.

Proof: As stated in Section II, there exist $\kappa_l, \kappa_h, \vartheta > 0$ such that for all $\boldsymbol{x} \in \mathbb{S}$, all eigenvalues of $\nabla_{\boldsymbol{x}}$ are within the interval (κ_l, κ_h) and $||\boldsymbol{b}_{\boldsymbol{x}}|| < \vartheta$. Hence, $\det \Delta_{\boldsymbol{x}}$ is within the interval $(1/\kappa_h^M, 1/\kappa_l^M)$. As a result

$$\begin{aligned}-\frac{\kappa_h}{2}||\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{x}}||^2 &\leq -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{x}})^{\intercal}\nabla_{\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{x}}) \\ &\leq -\frac{\kappa_l}{2}||\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{x}}||^2. \end{aligned} \tag{22}$$

Hence, for any $0<\epsilon<1$ and $||\boldsymbol{y}|| > \vartheta/\epsilon$, we have

$$\begin{aligned}-\frac{\kappa_h}{2}(1+\epsilon)^2||\boldsymbol{y}||^2 &\leq -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{x}})^{\intercal}\nabla_{\boldsymbol{x}}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{x}}) \\ &\leq -\frac{\kappa_l}{2}(1-\epsilon)^2||\boldsymbol{y}||^2 \end{aligned} \tag{23}$$

and

$$\begin{aligned}\frac{\exp(-\frac{\kappa_h}{2}(1+\epsilon)^2||\boldsymbol{y}||^2)}{\sqrt{(2\pi/\kappa_l)^M}} &\leq P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) \\ &\leq \frac{\exp(-\frac{\kappa_l}{2}(1-\epsilon)^2||\boldsymbol{y}||^2)}{\sqrt{(2\pi/\kappa_h)^M}}. \end{aligned} \tag{24}$$

Let $\beta' = \frac{\kappa_h}{2}(1+\epsilon)^2$ and $\beta = \frac{\kappa_l}{2}(1-\epsilon)^2$. Then $\beta', \beta > 0$. If α is large enough, then

$$\exp(-\alpha-\beta'||\boldsymbol{y}||^2) \leq P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) \leq \exp(\alpha-\beta||\boldsymbol{y}||^2)$$

for all $\boldsymbol{y} \in \mathbb{R}^M$ and the lemma is proved. □

Proposition 3 (Continuity): $H_{\boldsymbol{Y}}(\mu)$, $H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu)$, and $I(\mu)$ are continuous over $\Lambda_{\mathbb{S}}$.

Proof: Let $\{\mu_i\}_{i=0}^{\infty}$ be a sequence in $\Lambda_{\mathbb{S}}$ such that $\mu_i \Longrightarrow \mu_0$. For any fixed $\boldsymbol{y} \in \mathbb{R}^M$, $P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$ is a continuous function of $\boldsymbol{x}$ over $\mathbb{S}$. Hence,

$$\lim_{i\to\infty}\int P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})\,d\mu_i = \int P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})\,d\mu_0$$

i.e., $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_i)$ is pointwise convergent to $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)$. By Lemma 3, there exist $\alpha, \beta, \beta' > 0$ such that

$$|-P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_i)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_i)| \leq (\alpha+\beta'||\boldsymbol{y}||^2)\exp(\alpha-\beta||\boldsymbol{y}||^2).$$

By the Lebesgue convergence theorem

$$\begin{aligned}\lim_{i\to\infty} -\int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_i)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_i)\,d\boldsymbol{y} \\ = -\int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)\,d\boldsymbol{y}. \end{aligned} \tag{25}$$

Thus, $H_{\boldsymbol{Y}}(\mu)$ is continuous over $\Lambda_{\mathbb{S}}$.

The continuity of $H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu)$ follows from the fact that $\frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{x}}]$ is continuous over $\mathbb{S}$ and

$$H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu) \triangleq \int \frac{1}{2}\log[(2\pi e)^M \det \Delta_{\boldsymbol{x}}]\,d\mu.$$

Finally, as $I(\mu) = H_{\boldsymbol{Y}}(\mu) - H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu)$, the continuity of $I(\mu)$ is established. □

Lemma 4: Consider a constant CG channel $\boldsymbol{Y}=\boldsymbol{H}\boldsymbol{X}+\boldsymbol{Z}$. Let $\boldsymbol{R}=\boldsymbol{H}\boldsymbol{X}$ and $P_{\boldsymbol{R}}(\mu)$ be the probability measure of $\boldsymbol{R}$ where μ is the probability measure of $\boldsymbol{X}$. Then, $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)=P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_2)$ if and only if $P_{\boldsymbol{R}}(\mu_1)=P_{\boldsymbol{R}}(\mu_2)$. Consequently, if $\boldsymbol{H}$ is left-invertible, then $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)=P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_2)$ if and only if $\mu_1=\mu_2$.

Proof: It is obvious that $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)$ is equal to the convolution of $P_{\boldsymbol{Z}}(\boldsymbol{z})$ and $P_{\boldsymbol{R}}(\mu)$, where $P_{\boldsymbol{Z}}(\boldsymbol{z})$ is the probability density function of the Gaussian noise $\boldsymbol{Z}$. Since the Fourier transform of $P_{\boldsymbol{Z}}(\boldsymbol{z})$ is nonzero everywhere, $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)$ and $P_{\boldsymbol{R}}(\mu)$ are in one-to-one correspondence (see [1] and Lemma 7). In other words, for any input probability measures μ_1 and μ_2, $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)=P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_2)$ if and only if $P_{\boldsymbol{R}}(\mu_1)=P_{\boldsymbol{R}}(\mu_2)$. □

Definition 6: Let $J(\mu)$ be a real-valued function defined on $\Lambda_{\mathbb{S}}$ and $\mu_0 \in \Lambda_{\mathbb{S}}$. If the limit

$$J'_{\mu_0}(\mu_1) = \lim_{\theta\to 0^+}\frac{J((1-\theta)\mu_0+\theta\mu_1)-J(\mu_0)}{\theta} \tag{26}$$

exists for all $\mu_1 \in \Lambda_{\mathbb{S}}$, then the function $J(\mu)$ is called weakly differentiable [1] at μ_0, and the function $J'_{\mu_0}(\mu_1)$ is its weak derivative at μ_0.

Note that, if $J(\mu)$ is concave and is weakly differentiable on $\Lambda_{\mathbb{S}}$, then $J(\mu_o)$ is maximized if and only if $J'_{\mu_0}(\mu_1) \leq 0$ for all $\mu_1 \in \Lambda_{\mathbb{S}}$.

Proposition 4 (Weak Differentiability): The mutual information function $I(\mu)$ is weakly differentiable in $\Lambda_{\mathbb{S}}$, and

$$I'_{\mu_0}(\mu_1) = \int Q(\boldsymbol{x};\mu_0)d\mu_1 - H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu_1) - I(\mu_0).$$

Proof: Let $0 \le \theta \le 1$ and $\mu_\theta = (1-\theta)\mu_0 + \theta\mu_1$. It can be shown easily that $\frac{I(\mu_\theta)-I(\mu_0)}{\theta}$ is equal to

$$-I(\mu_0) - H_{\boldsymbol{Y}|\boldsymbol{X}}(\mu_1) - \int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)d\boldsymbol{y} + \frac{1-\theta}{\theta}\int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)\log\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)}{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)}d\boldsymbol{y}. \quad (27)$$

Since $\int Q(\boldsymbol{x};\mu_0)d\mu_1 = -\int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)d\boldsymbol{y}$, to prove Proposition 4, it suffices to prove the following claims.

Claim 1:

$$\lim_{\theta\to 0^+}\int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)d\boldsymbol{y} = \int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)d\boldsymbol{y}.$$

Claim 2:

$$\lim_{\theta\to 0^+}\frac{1}{\theta}\int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)\log\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)}{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)}d\boldsymbol{y} = 0.$$

As the limit $\lim_{\theta\to 0^+} P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)$ is equal to $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)$ and there exists $\alpha,\beta,\beta' > 0$ such that

$$|P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)| \le (\alpha+\beta'||\boldsymbol{y}||^2)\exp(\alpha-\beta||\boldsymbol{y}||^2)$$

the first claim thus follows from the Lebesgue convergence theorem.

To prove the second claim, first notice that for $c, c_1, c_2 > 0$, $\log(1-\theta+\theta c) \ge \theta\log c$ and $c_1\log(c_2/c_1) \le c_2 - c_1$. Let

$$c = P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)/P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)$$
$$c_1 = P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)$$

and

$$c_2 = P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta).$$

Then it is easy to show that

$$P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)\log\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)}{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)} \le \frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)}{\theta}\log\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)}{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)} \le P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1) - P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0).$$

By L'Hospital's rule, it can be shown that

$$\lim_{\theta\to 0^+}\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)}{\theta}\log\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)}{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)} = P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1) - P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0).$$

Since $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)\log\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1)}{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)}$ and $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1) - P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)$ are integrable, by the Lebesgue convergence theorem

$$\lim_{\theta\to 0^+}\int\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)}{\theta}\log\frac{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_\theta)}{P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)}\,d\boldsymbol{y} = \int P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_1) - P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_0)\,d\boldsymbol{y} \quad (28)$$
$$= 0. \quad (29)$$

The proposition is thus proved. □

Lemma 5: $Q(\boldsymbol{x};\mu_o)$ is a continuous function of $\boldsymbol{x}$ over $\mathbb{S}$.

Proof: Suppose $\{\boldsymbol{x}^{(i)}\}_{i=1}^{\infty}$ is a convergent sequence in $\mathbb{S}$ with limit $\boldsymbol{x}^{(0)}$. For any $\boldsymbol{y}\in\mathbb{R}^M$, the sequence

$$\{-P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}^{(i)})\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o)\}_{i=1}^{\infty}$$

is pointwise convergent to $-P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}^{(0)})\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o)$. Again, by the Lebesgue convergence theorem, we have

$$\lim_{i\to\infty} -\int P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}^{(i)})\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o)\,d\boldsymbol{y} = -\int P_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}^{(0)})\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o)\,d\boldsymbol{y} \quad (30)$$

and the result follows. □

Proposition 5: Let $d(\boldsymbol{w})$ be defined as in (11). Then $d(\boldsymbol{w})$ is holomorphic over $\mathbb{A}(\Delta_{\boldsymbol{w}})$.

Proof: Suppose $\mathbb{V}$ is a closed and bounded subset of $\mathbb{A}(\Delta_{\boldsymbol{w}})$. As the entries in $\boldsymbol{b}_{\boldsymbol{w}}$ and $\nabla_{\boldsymbol{w}}$ are continuous functions of $\boldsymbol{w}$, there exists $\widetilde{\vartheta}, \widetilde{\kappa_h}, \widetilde{\kappa_l}, \widetilde{\varsigma_h}, \widetilde{\varsigma_l}, \widetilde{\varsigma} > 0$ such that for all $\boldsymbol{w}\in\mathbb{V}$, we have 1) $||\boldsymbol{b}_{\boldsymbol{w}}|| \le \widetilde{\vartheta}$, 2) the eigenvalues of the matrix $\mathrm{re}\nabla_{\boldsymbol{w}}$ are within the interval $(\widetilde{\kappa_l}, \widetilde{\kappa_h})$, 3) $||\ \mathrm{im}(\nabla_{\boldsymbol{w}})\ \mathrm{im}(\boldsymbol{b}_{\boldsymbol{w}})\ || < \widetilde{\varsigma}$, and 4) the magnitude of $\det\nabla_{\boldsymbol{w}}$ is within the interval $(\widetilde{\varsigma_l}, \widetilde{\varsigma_h})$.

For any scalar ϵ and vector $\boldsymbol{y}\in\mathbb{R}^M$ such that $1 > \epsilon > 0$ and $||\boldsymbol{y}|| > \max(\widetilde{\vartheta}/\epsilon, (1+\epsilon)\widetilde{\varsigma}/\epsilon)$, we have

$$\mathrm{re}\left((\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})^{\intercal}\nabla_{\boldsymbol{w}}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})\right) \quad (31)$$
$$= \mathrm{re}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})^{\intercal}\mathrm{re}(\nabla_{\boldsymbol{w}})\mathrm{re}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}}) - \mathrm{im}(\boldsymbol{b}_{\boldsymbol{w}})^{\intercal}\mathrm{re}(\nabla_{\boldsymbol{w}})\mathrm{im}(\boldsymbol{b}_{\boldsymbol{w}}) - 2\mathrm{re}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})^{\intercal}\mathrm{im}(\nabla_{\boldsymbol{w}})\mathrm{im}(\boldsymbol{b}_{\boldsymbol{w}}) \quad (32)$$
$$\ge \widetilde{\kappa_l}||\mathrm{re}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})||^2 - \widetilde{\kappa_h}||\mathrm{im}(\boldsymbol{b}_{\boldsymbol{w}})||^2 - 2||\mathrm{re}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})||\widetilde{\varsigma} \quad (33)$$
$$\ge ||\boldsymbol{y}||^2\left[\widetilde{\kappa_l}(1-\epsilon)^2 - \widetilde{\kappa_h}\epsilon^2 - 2\epsilon\right]. \quad (34)$$

Let $\widetilde{\beta}$ be $(\widetilde{\kappa_l}(1-\epsilon)^2 - \widetilde{\kappa_h}\epsilon^2 - 2\epsilon)/2$. By choosing ϵ small enough, $\widetilde{\beta}$ is positive. Let $q(\boldsymbol{y};\boldsymbol{w})$ be

$$\exp(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}})^{\intercal}\nabla_{\boldsymbol{w}}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}}))/\sqrt{(2\pi)^M\det\Delta_{\boldsymbol{w}}}$$

and $f(\boldsymbol{y},\boldsymbol{w})$ be $-q(\boldsymbol{y},\boldsymbol{w})\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o)$. Thus, for all

$$||\boldsymbol{y}|| > \max(\widetilde{\vartheta}/\epsilon, (1+\epsilon)\widetilde{\varsigma}/\epsilon)$$

we have

$$|q(\boldsymbol{y};\boldsymbol{w})| \le \exp(-\widetilde{\beta}||\boldsymbol{y}||^2)/\sqrt{(2\pi)^M/\widetilde{\varsigma_h}}.$$

As a result, by picking a positive number $\widetilde{\alpha}$ large enough, we have

$$|q(\boldsymbol{y};\boldsymbol{w})| \le \exp(\widetilde{\alpha}-\widetilde{\beta}||\boldsymbol{y}||^2), \quad \text{for all } \boldsymbol{y}\in\mathbb{R}^M$$

and

$$|f(\boldsymbol{y},\boldsymbol{w})| \le (\alpha+\beta'||\boldsymbol{y}||^2)\exp(\widetilde{\alpha}-\widetilde{\beta}||\boldsymbol{y}||^2).$$

As a corollary, the integral $\int|f(\boldsymbol{y},\boldsymbol{w})|\,d\boldsymbol{y}$ is uniformly convergent over $\mathbb{V}$.

On the other hand, since the entries in $\Delta_{\boldsymbol{w}}$ and $\boldsymbol{b}_{\boldsymbol{w}}$ are well behaved and $\mathrm{re}(\det\Delta_{\boldsymbol{w}}) > 0$ for all $\boldsymbol{w}\in\mathbb{A}(\Delta_{\boldsymbol{w}})$, the function $f(\boldsymbol{y};\boldsymbol{w})$ is a holomorphic function over $\mathbb{A}(\Delta_{\boldsymbol{w}})$ for any $\boldsymbol{y}\in\mathbb{R}^M$. Therefore, according to the differentiation lemma [24], $d(\boldsymbol{w})$ is holomorphic over $\mathbb{A}(\Delta_{\boldsymbol{w}})$. □

Lemma 6: Let $d(\boldsymbol{w})$ be defined as in (11). If $\boldsymbol{w}^* \in \mathbb{A}(\Delta_{\boldsymbol{w}})$ such that $\nabla_{\boldsymbol{w}^*}$ and $\boldsymbol{b}_{\boldsymbol{w}^*}$ are real, then for some $\alpha \geq 0$ and $\beta \geq 0$, we have $d(\boldsymbol{w}^*) \geq -\alpha + \beta \mathrm{tr}[\Delta_{\boldsymbol{w}^*}]$.

Proof: Since $\nabla_{\boldsymbol{w}^*}$ is real and $\boldsymbol{w}^* \in \mathbb{A}(\Delta_{\boldsymbol{w}})$, the matrix $\Delta_{\boldsymbol{w}^*}$ is symmetric and positive definite. By Lemma 3, we have $-\log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o) \geq -\alpha + \beta\|\boldsymbol{y}\|^2$. Let

$$C(\boldsymbol{y},\boldsymbol{w}) = \frac{\exp(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}^*})^{\mathsf{T}}\nabla_{\boldsymbol{w}^*}(\boldsymbol{y}-\boldsymbol{b}_{\boldsymbol{w}^*}))}{\sqrt{(2\pi)^M \det \Delta_{\boldsymbol{w}^*}}}. \tag{35}$$

Then, we have

$$\begin{aligned} d(\boldsymbol{w}^*) &= -\int C(\boldsymbol{y},\boldsymbol{w}) \log P_{\boldsymbol{Y}}(\boldsymbol{y};\mu_o) d\boldsymbol{y} && (36)\\ &\geq \int C(\boldsymbol{y},\boldsymbol{w})(-\alpha+\beta\|\boldsymbol{y}\|^2) d\boldsymbol{y} && (37)\\ &= -\alpha + \beta(\mathrm{tr}[\Delta_{\boldsymbol{w}^*}] + \|\boldsymbol{b}_{\boldsymbol{w}^*}\|^2) && (38)\\ &\geq -\alpha + \beta \mathrm{tr}[\Delta_{\boldsymbol{w}^*}]. && (39) \end{aligned}$$

□

Lemma 7: Let $\psi(\boldsymbol{y})$ be a Schwartz function [21] defined on $\mathbb{R}^M$ such that its Fourier transform $\hat{\psi}(\boldsymbol{y})$ is nonzero for all $\boldsymbol{y} \in \mathbb{R}^M$. Suppose Ψ is a tempered distribution (i.e., a linear and continuous functional defined on the set of Schwartz functions) and $\Psi * \psi$, the convolution between Ψ and ψ, is the zero tempered distribution. Then Ψ and its Fourier transform $\hat{\Psi}$ are the zero tempered distribution. In addition, if there exist two probability measures μ_0 and μ_1 such that $\Psi = \mu_1 - \mu_0$, then $\mu_1 = \mu_0$.

Proof: Let $\mathcal{D}$ be the set of Schwartz functions which have a compact support and ϕ_0 be a function in $\mathcal{D}$. As $\hat{\psi}(\boldsymbol{y})$ is nonzero for all $\boldsymbol{y} \in \mathbb{R}^M$, the function $\phi_1 \triangleq \phi_0/\hat{\psi}$ (i.e., $\phi_1(\boldsymbol{y}) = \phi_0(\boldsymbol{y})/\hat{\psi}(\boldsymbol{y})$ for all $\boldsymbol{y} \in \mathbb{R}^M$) is also in $\mathcal{D}$.

Since $(\Psi * \psi)$ is the zero tempered distribution, its Fourier transform $\hat{\Psi} \cdot \hat{\psi}$ is again the zero tempered distribution. Consequently, we have[4]

$$\begin{aligned} 0 &= \langle \phi_1, \hat{\Psi} \cdot \hat{\psi} \rangle && (40)\\ &= \langle \phi_0/\hat{\psi}, \hat{\Psi} \cdot \hat{\psi} \rangle && (41)\\ &= \langle \phi_0, \hat{\Psi} \rangle. && (42) \end{aligned}$$

As $\mathcal{D}$ is dense in the set of Schwartz functions[22], $\langle \phi, \hat{\Psi} \rangle = 0$ for any Schwartz function ϕ. In other words, $\hat{\Psi}$, and consequently, Ψ, are the zero tempered distribution.

Finally, if $\Psi = \mu_1 - \mu_0$ for some probability measures μ_0 and μ_1, then $\langle \phi, \mu_1 \rangle = \langle \phi, \mu_0 \rangle$ for any $\phi \in \mathcal{D}$. By Riesz representation theorem [25], $\mu_1 = \mu_0$. □

Corollary 9: Let $\boldsymbol{Z}$ be a Gaussian random variable and $P_{\boldsymbol{Z}}(\boldsymbol{z})$ be the corresponding probability density function. Suppose $g(\boldsymbol{z})$ is a continuous function such that $|g(\boldsymbol{z})| \leq \alpha + \beta\|\boldsymbol{z}\|^2$ for some $\alpha, \beta > 0$. If $P_{\boldsymbol{Z}}(\boldsymbol{z}) * g(\boldsymbol{z})$ is the zero function, then $g(\boldsymbol{z})$ is also the zero function.

Proof: It can be proved easily that 1) $P_{\boldsymbol{Z}}(\boldsymbol{z})$ is a Schwartz function such that its Fourier transform is nonzero everywhere, and 2) $g(\boldsymbol{z})$ is a tempered distribution. Therefore, by Lemma 7, $g(\boldsymbol{z})$ is the zero tempered distribution, which further implies that $g(\boldsymbol{z})$ is the zero function. □

Lemma 8: Consider a constant CG channel $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{Z}$. Let $\boldsymbol{X}$ be Gaussian distributed with probability measure ϕ. Then we can construct a sequence of probability measures $\{\nu_i\}_{i=1}^{\infty}$ such that 1) $\mathbb{E}_{\nu_i}$ is closed and bounded, 2) $\boldsymbol{E}_{\nu_i}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}] \leq \boldsymbol{E}_{\phi}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}]$, 3) $\nu_i \Longrightarrow \phi$ in the *Lévy* metric, and 4) $\lim_{i\to\infty} H_{\boldsymbol{Y}}(\nu_i) = H_{\boldsymbol{Y}}(\phi)$, or equivalently, $\lim_{i\to\infty} I(\nu_i) = I(\phi)$.

Proof: Assume without loss of generality that $\boldsymbol{E}_{\phi}[\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}]$ is positive definite. Hence, ϕ is indeed characterized by a probability density function, which is denoted by $\phi(\boldsymbol{x})$ for simplicity. For any positive integer i, let $\mathbb{V}_i = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \leq i\}$, and $\nu_i(x)$ be the following "truncated" Gaussian probability density function

$$\nu_i(\boldsymbol{x}) = \begin{cases} \alpha_i \phi(\boldsymbol{x}), & \text{if } \boldsymbol{x} \in \mathbb{V}_i \\ 0, & \text{otherwise} \end{cases} \tag{43}$$

where $\alpha_i > 1$ is a normalizing constant. Clearly, conditions 1) and 2) are satisfied, and $\lim_{i\to\infty} \alpha_i = 1$. For any continuous and bounded function $f(\boldsymbol{x})$ defined on $\mathbb{R}^N$, it is apparent that

$$\lim_{i\to\infty} \int f(\boldsymbol{x})\nu_i(\boldsymbol{x})d\boldsymbol{x} = \int f(\boldsymbol{x})\phi(\boldsymbol{x})d\boldsymbol{x}.$$

Hence, by Lemma 2, $\nu_i \Longrightarrow \phi$. It remains to prove that $\lim_{i\to\infty} H_{\boldsymbol{Y}}(\nu_i) = H_{\boldsymbol{Y}}(\phi)$.

Let $b_i(\boldsymbol{y}) = P_{\boldsymbol{Y}}(\boldsymbol{y};\nu_i)/\alpha_i$. Since $\nu_i \Longrightarrow \phi$

$$\lim_{i\to\infty} b_i(\boldsymbol{y}) = P_{\boldsymbol{Y}}(\boldsymbol{y};\phi).$$

Also, as $\nu_i(\boldsymbol{x})/\alpha_i \leq \phi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^N$, $b_i(\boldsymbol{y}) \leq P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)$ for all $\boldsymbol{y} \in \mathbb{R}^M$.

Let $\vartheta > 0$ such that $P_{\boldsymbol{Y}}(\boldsymbol{y};\phi) < 1/e$ for all $\|\boldsymbol{y}\| \geq \vartheta$. As $-t\log t$ is increasing and nonnegative on the interval $[0, 1/e]$, $|-b_i(\boldsymbol{y})\log b_i(\boldsymbol{y})| \leq |-P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)|$ for all $\|\boldsymbol{y}\| \geq \vartheta$. By the Lebesgue convergence theorem

$$\begin{aligned} \lim_{i\to\infty} \int_{\|\boldsymbol{y}\|\geq\vartheta} & -b_i(\boldsymbol{y})\log b_i(\boldsymbol{y}) d\boldsymbol{y} \\ &= \int_{\|\boldsymbol{y}\|\geq\vartheta} -P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\phi) d\boldsymbol{y}. \end{aligned}$$

Similarly, as the function $|-b_i(\boldsymbol{y})\log b_i(\boldsymbol{y})|$ is bounded for $\|\boldsymbol{y}\| \leq \vartheta$, the limit

$$\lim_{i\to\infty} \int_{\|\boldsymbol{y}\|\leq\vartheta} -b_i(\boldsymbol{y})\log b_i(\boldsymbol{y}) d\boldsymbol{y}$$

is equal to $\int_{\|\boldsymbol{y}\|\leq\vartheta} -P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)d\boldsymbol{y}$. Consequently, as i goes to infinity, $-\int b_i(\boldsymbol{y})\log b_i(\boldsymbol{y})d\boldsymbol{y}$ converges to

$$-\int P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)d\boldsymbol{y}$$

and, as a result, the limit

$$\lim_{i\to\infty} -\int P_{\boldsymbol{Y}}(\boldsymbol{y};\nu_i)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\nu_i)d\boldsymbol{y}$$

is also equal to

$$-\int P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)\log P_{\boldsymbol{Y}}(\boldsymbol{y};\phi)d\boldsymbol{y}$$

i.e., $\lim_{i\to\infty} H_{\boldsymbol{Y}}(\nu_i) = H_{\boldsymbol{Y}}(\phi)$ and the lemma is proved. □

Proposition 6: Consider a sequence of constant CG channels $\{(\Delta, \boldsymbol{H}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{S}_j)\}_{j=1}^{\infty}$ such that for any closed and bounded subset $\mathbb{V}$ of $\mathbb{R}^N$, $\mathbb{V} \subseteq \mathbb{S}_j$ for sufficiently large j.

[4]For any Schwartz function ψ and tempered distribution Ψ, we denote the function value of Ψ at ψ by $\langle \phi, \Psi \rangle$.

Let μ_j and Φ be the capacity-achieving measures for the channels $(\Delta, \boldsymbol{H}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{S}_j)$ and $(\Delta, \boldsymbol{H}\boldsymbol{x}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{R}^N)$, respectively. Then $\lim_{j\to\infty} I(\mu_j) = I(\Phi)$, or equivalently, $\lim_{j\to\infty} H_{\boldsymbol{Y}}(\mu_j) = H_{\boldsymbol{Y}}(\Phi)$.

Proof: By Lemma 8, we can construct a sequence $\{\nu_i\}_{i=1}^{\infty}$ such that 1) $\mathbb{E}_{\nu_i}$ is closed and bounded, 2) $\boldsymbol{E}_{\nu_i}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}] \le \boldsymbol{E}_{\Phi}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}]$, and 3) $\lim_{i\to\infty} I(\nu_i) = I(\Phi)$. It is obvious that for sufficiently large j, $\mathbb{E}_{\nu_i} \subseteq \mathbb{S}_j$. Hence, $I(\nu_i) \le I(\mu_j) \le I(\Phi)$, and thus, the proposition follows. □

Consider a constant CG channel with a diagonal square channel matrix $\boldsymbol{D}$ with diagonal entries χ_n. Let Φ be the capacity-acheiving probability measure for the channel $(\boldsymbol{I}_N, \boldsymbol{D}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{R}^N)$, and $P_n \triangleq \boldsymbol{E}_{\Phi}[X_n^2]$ for $n \in \mathsf{N}$.

Lemma 9: Let μ be an input probability measure such that $\boldsymbol{E}_{\mu}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}] \le \gamma$. For the above channel $(\boldsymbol{I}_N, \boldsymbol{D}, \{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} - \gamma\}, \mathbb{R}^N)$, we have

$$D(P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)\|P_{\boldsymbol{Y}}(\boldsymbol{y};\Phi)) \le H_{\boldsymbol{Y}}(\Phi) - H_{\boldsymbol{Y}}(\mu)$$

where $D(P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)\|P_{\boldsymbol{Y}}(\boldsymbol{y};\Phi))$ is the Kullback–Leibler distance between $P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)$ and $P_{\boldsymbol{Y}}(\boldsymbol{y};\Phi)$. Consequently

$$\int |P_{\boldsymbol{Y}}(\boldsymbol{y};\mu) - P_{\boldsymbol{Y}}(\boldsymbol{y};\Phi)| d\boldsymbol{y} \le \sqrt{2H_{\boldsymbol{Y}}(\Phi) - 2H_{\boldsymbol{Y}}(\mu)}$$

according to Pinsker's inequality [19].

Proof: According to the water-filling algorithm, there exists $\vartheta > 0$ such that $\chi_n^2/(\chi_n^2 P_n + 1) \le 1/\vartheta$ and $P_n\chi_n^2/(\chi_n^2 P_n + 1) = P_n/\vartheta$ for all $n \in \mathsf{N}$. Let Δ^* be the covariance matrix of $\boldsymbol{Y}$ when the input probability measure is Φ. Let $C = -H_{\boldsymbol{Y}}(\mu) + \frac{1}{2}\log[(2\pi)^N \det \Delta^*]$. We have

$$\begin{aligned} &D(P_{\boldsymbol{Y}}(\boldsymbol{y};\mu)\|P_{\boldsymbol{Y}}(\boldsymbol{y};\Phi)) \\ &\quad = C + \frac{1}{2}\sum_{n\in\mathsf{N}} \frac{\chi_n^2 \boldsymbol{E}_{\mu}[X_n^2] + 1}{\chi_n^2 P_n + 1} && (44) \\ &\quad \le C + \frac{1}{2}\sum_{n\in\mathsf{N}} \frac{\boldsymbol{E}_{\mu}[X_n^2]}{\vartheta} + \frac{1}{2}\sum_{n\in\mathsf{N}} \frac{1}{\chi_n^2 P_n + 1} && (45) \\ &\quad \le C + \frac{\gamma}{2\vartheta} + \frac{1}{2}\sum_{n\in\mathsf{N}} \frac{1}{\chi_n^2 P_n + 1} && (46) \\ &\quad = H_{\boldsymbol{Y}}(\Phi) - H_{\boldsymbol{Y}}(\mu). && (47) \end{aligned}$$

□

Lemma 10: Let Γ be the set of all input probability measures μ of $\boldsymbol{X}$ such that $\boldsymbol{E}_{\mu}[\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}] \le \gamma$. Then Γ is sequentially compact.

Proof: For any $\epsilon > 0$, let $\mathbb{K}$ be the closed and bounded set $\{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} \le \gamma/\epsilon\}$. Hence, if $\mu \in \Gamma$, then $\mu(\mathbb{K}) \ge 1 - \epsilon$. Therefore, Γ is tight. Let $\{\mu_i\}_{i=1}^{\infty}$ be a convergent sequence in Γ such that it converges to μ_0. By Lemma 2, we have

$$\int \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} d\mu_0 \le \underline{\lim}_{i\to\infty} \int \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} d\mu_i \le \gamma.$$

Hence, $\mu_0 \in \Gamma$, and thus, Γ is closed. The sequential compactness of Γ then follows from Lemma 2. □

ACKNOWLEDGMENT

The authors would like to thank the referees and the Associate Editor for their detailed comments and suggestions, which have significantly improved the readability of the paper.

REFERENCES

[1] J. G. Smith, "The information capacity of amplitude and variance-constrained scalar Gaussian channels," *Inf. Contr.*, vol. 18, pp. 203–219, 1971.

[2] S. Shamai (Shitz), "Capacity of a pulse amplitude modulated direct detection photon channel," *Proc. Inst. Elec. Eng.*, pt. 1, vol. 137, no. 6, pp. 424–430, Dec. 1990.

[3] S. Shamai (Shitz) 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, Jul. 1995.

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

[5] M. Katz and S. Shamai (Shitz), "On the capacity-achieving distribution of the discrete-time noncoherent additive white noise channel," in *Proc. 2002 IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, Jun./Jul. 2002, p. 165.

[6] W. Oettli, "The capacity-achieving input distribution for some amplitude limited channels with additive noise," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 3, pp. 372–374, May 1974.

[7] A. Das, "Capacity-achieving distributions for non-Gaussian additive noise channels," in *Proc. 2000 IEEE Int. Symp. Information Theory*, Sorrento, Italy, Jun. 2000, p. 432.

[8] R. Palanki, "On the capacity achieving distributions of some fading channels," in *Proc. 40th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2002, pp. 337–346.

[9] M. C. Gursoy, H. V. Poor, and S. Verdú, "On the capacity-achieving distribution of the noncoherent Rician fading channel," in *Proc. 2003 Canadian Workshop on Information Theory*, Waterloo, ON, Canada, May 2003, pp. 24–27.

[10] R. Nuriyev and A. Anastasopoulos, "Capacity characterization for the noncoherent block independent AWGN channel," in *Proc. 2003 IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul. 2003, p. 373.

[11] P. Hou, B. J. Belzer, and T. R. Fischer, "On the capacity of the paritially coherent additive white Gaussian noise channel," in *Proc. 2003 IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul. 2003, p. 372.

[12] J. Huang and S. Meyn, "Characterization and computation of optimal distributions for channel coding," *IEEE Trans. Inf. Theory*, submitted for publication.

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

[14] S. Hranilovic and F. R. Kschischang, "Optical intensity-modulated direct detection channels: Signal space and Lattice codes," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1385–1399, Jun. 2003.

[15] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh fading," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.

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

[17] R. C. Gunning, *Introduction to Holomorphic Functions of Several Variables*. Pacific Grove, CA: Wadsworth and Brooks Cole, 1990.

[18] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.

[20] R. M. Range, *Holomorphic Functions and Integral Representations in Several Complex Variables*. New York: Springer-Verlag, 1986.

[21] W. Rudin, *Functional Analysis*. New York: McGraw-Hill, 1973.

[22] R. S. Pathak, *A Course in Distribution Theory and Applications*. Boca Raton, FL: CRC Press/Narosa, 2000.

[23] D. W. Stroock, *Probability Theory, an Analytic View*. New York: Cambridge Univ. Press, 1993.

[24] S. Lang, *Complex Analysis*. New York: Springer-Verlag, 1999.

[25] G. B. Folland, *Real Analysis : Modern Techniques and Their Applications*. New York: Wiley, 1984.

[26] G. N. Iyengar, "Voice channel," in *Proc. 31st Asilomar Conf. Signals, Systems and Computers*, vol. 2, Pacific Grove, CA, 1997, pp. 1354–1358.

[27] A. Lapidoth and S. Shamai (Shitz), "The Poisson multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 488–501, Mar. 1998.

[28] G. B. Folland, *Fourier Analysis and its Applications*. Pacific Grove, CA: Wadsworth and Brooks Cole Advanced Books and Software, 1992.