Characterization and Computation of Optimal Distributions for Channel Coding

Jianyi Huang and Sean Meyn*

September 17, 2003

Abstract

This paper concerns the structure of optimal codes for stochastic channel models. An investigation of an associated dual convex program reveals that the optimal distribution in channel coding is typically discrete. Based on this observation we obtain the following theoretical conclusions, as well as new algorithms for constructing efficient channel codes:

(i) Under general conditions, for low SNR the optimal random code is defined by a distribution whose magnitude is binary.

(ii) Simple discrete approximations may be highly accurate even in cases where the optimal distribution is known to be absolutely continuous with respect to Lebesgue measure.

(iii) A new class of algorithms is introduced, based on the cutting-plane method, to generate discrete distributions that are optimal within a prescribed class.

Keywords: Information theory; channel coding; fading channels.

*Department of Electrical and Computer Engineering, the Coordinated Science Laboratory, and the University of Illinois, 1308 W. Main Street, Urbana, IL 61801, URL `http://black.csl.uiuc.edu:80/∼meyn` (s-meyn@uiuc.edu). Work supported in part by the National Science Foundation through ITR 00-85929

1 Introduction

Since Shannon's celebrated paper [29], published in 1948, channel capacity has become a fundamental topic in information theory. The i.i.d. additive white Gaussian noise (AWGN) channel has been the focus of most research due to its tractability, and because this model reflects the behavior of many communication channels. It is well known that the optimal input distribution is i.i.d. Gaussian in this special case (see e.g. [12, 15, 8]).

More recently, there has been a significant research effort on fading channels, such as found in wireless communication systems. Early papers restricted to Gaussian dispersive channels [26, 15, 25], while more recent papers study a range of complex fading models with correspondingly complex analysis. A recent survey is contained in [4].

A theme in recent work is the discovery of an increasing list of special cases in which the magnitude of the optimal input distribution is *discrete*, with *finite* support. Examples include,

(i) *Amplitude and variance constrained Gaussian channels* [30]:

If the input is not only constrained by the average power but also limited by a given peak power constraint, the optimal input distribution is shown to be achieved by a unique *discrete* random variable taking on a *finite* number of values.

(ii) *Average and peak power limited quadrature Gaussian channels* [27]:

In this paper the conclusions of [30] are generalized to the complex case. The magnitude of the capacity-achieving distribution is shown to be *discrete*, with *finite* support. Moreover, its phase is uniformly distributed, and independent of the magnitude.

(iii) *Discrete-time Rayleigh fading channels* [1]:

The capacity-achieving distribution is shown to be *discrete* in magnitude with a *finite* number of mass points, one of them located at the origin. The exact number and location of these mass points vary significantly with SNR. Extensions of this result have appeared recently in [10].

(iv) *Single antenna Rayleigh block-fading channels* [22]:

It is proved that for a single-antenna block-fading channel, the capacity achieving random variable is the product of a *discrete* real random variable and an *isotropically* distributed unit vector.

(v) *Mobile multiple-antenna communication link in a Rayleigh flat fading* [20]:

The capacity of a multiple antenna (M transmitters, N receivers) block-fading Rayleigh channel, in which the propagation coefficient is assumed to be constant over any T-symbol interval, is achieved when the $T \times M$ transmitted signal matrix is equal to the product of two statistically independent matrices: a $T \times T$ *isotropically* distributed unitary matrix, times a certain $T \times M$ random matrix that is *diagonal, real*, *nonnegative* and *discrete*.

It is conjectured in [22] that optimal distributions have finite support in many other fading models. In Section 3 we provide several propositions and examples to support this principle.

In conclusion, the AWGN channel is a very special case, in that optimal distributions rarely possess a density. Moreover, we show in examples below that even for the AWGN channel in which the optimal distribution is continuous, there exist very simple discrete distributions that very nearly achieve capacity. For the non-coherent Rayleigh fading channel, Gaussian input is shown to generate bounded mutual information as SNR goes to infinity [10].

We note that discrete distributions also arise in a worst-case analysis of many statistical models. In particular, for an additive-noise communication channel with fixed binary input, the worst-case noise distribution is supported on an integer lattice [28]. A general framework is developed in [24, 23] in the analysis of admission control algorithms, and general robust hypothesis testing problems. A key observation is that the worst-case distribution is always discrete.

A second, perhaps more important issue is computation. It is far easier to establish qualitative properties of the optimal distribution than to obtain a closed form expression. This is probably impossible in all but the simplest models. From these observations we are led to the following question: *Given a channel model, can we find a simple, discrete distribution that almost achieves the optimal mutual information?* If so, then there is no need to exactly optimize.

To obtain discrete approximations to the optimal input distribution we approximate the concave mutual information functional by a *piecewise linear functional.* Optimization of this approximation may be cast as an infinite-dimensional linear program, and the optimizer may be taken as a basic feasible solution, or extreme point in the constraint set. Such extreme points are in fact discrete distributions, and the number of support points grows at most linearly with the number of linear functions used in the approximations. In Section 4 we construct an algorithm of this form based on the cutting-plane algorithm.

These results were previously published in abridged form in [17].

The remainder of the paper is organized as follows. In Section 2 we review the Kuhn-Tucker theory for channel coding. In Section 3 we explain why the optimal input distribution is discrete in many channel models, and in Section 4 we present theory and numerical results for the cutting-plane algorithm. Section 5 contains conclusions and topics of future research.

2 Capacity of Memoryless Channels

2.1 Models

We consider in this paper a stationary, memoryless channel with input alphabet X, output alphabet Y, and transition density defined by

$$\mathsf{P}(Y \in dy \mid X = x) = p(y|x)\, dy\,, \qquad x \in \mathsf{X},\ y \in \mathsf{Y}.$$

Throughout the paper it is assumed that Y is equal to either $\mathbb{R}$ or $\mathbb{C}$, and we assume that X is a closed subset of $\mathbb{R}$. Channel models in which X is equal to $\mathbb{C}$ will be reduced to this form, as described below.

Let $\mathcal{M}$ denote the set of probability measures on $\mathcal{B}(\mathsf{X})$. For a given input distribution $\mu \in \mathcal{M}$, we denote by $p(y|\mu)$ the resulting output distribution given by $p(y|\mu) := \int p(y|x)\,\mu(dx)$. When $\mu = \delta_x$ we simplify notation by setting $p(y|\delta_x) = p(y|x)$.

Throughout the paper we restrict to non-coherent channels in which neither the sender nor the receiver knows the channel state. In this case, channel capacity is determined by the mutual information which is defined as the functional

$$I(\mu) = \iint \ln\Big(\frac{p(y|x)}{p(y|\mu)}\Big)\,\mu(dx)p(y|x)\,dy\,, \qquad \mu \in \mathcal{M}. \tag{1}$$

The *channel sensitivity function* and *channel discrimination function* are defined, respectively, by

$$g_\mu(x) \;:=\; D(p(\,\cdot\,|x)\|p(\,\cdot\,|\mu)) = \int \ln\big[p(y|x)/p(y|\mu)\big]\,p(y|x)\,dy\,, \tag{2}$$

$$g_0(x) \;:=\; g_{\delta_0}(x) = D(p(\,\cdot\,|x)\|p(\,\cdot\,|0))\,, \qquad \mu \in \mathcal{M}\,,\; x \in \mathsf{X}\,. \tag{3}$$

The mutual information may be expressed using this notation as $I(\mu) = \langle \mu, g_\mu\rangle$.

The magnitude of $g_0(x)$ indicates how easily an input $X = x$ may be discriminated against $X = 0$. Observe that $g_0(x) \geq 0$ for all $x \in \mathsf{X}$, and $g_0(0) = 0$. In Theorem 3.3 we find that the channel discrimination function is particularly valuable in analysis of the channel when the SNR is low.

A representation of mutual information that emphasizes its concave nature is given in the following proposition. The formula (4) will serve as a basis for the algorithms introduced in Section 4.

Proposition 2.1 *For any given any $\mu^\circ \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$,*

$$I(\mu^\circ) = \langle \mu^\circ, g_{\mu^\circ}\rangle = \min_{\mu\in\mathcal{M}} \langle \mu^\circ, g_\mu\rangle. \tag{4}$$

Proof We have for all $\mu, \mu^\circ \in \mathcal{M}$,

$$\langle \mu^\circ, g_\mu\rangle = I(\mu^\circ) + D\big(p(\,\cdot\,|\mu^\circ\|p(\,\cdot\,|\mu)\big) \geq I(\mu)\,.$$

By definition, this lower bound is attained with $\mu = \mu^\circ$. ☐

In this paper our main concern is the structure of distributions maximizing mutual information, subject to two linear constraints:

(i) The *average power constraint* that

$$\langle \mu, \phi\rangle \leq \sigma_P^2$$

where $\langle \mu, \phi\rangle := \int \phi(x)\;\mu(dx)$, and $\phi(x) := x^2$ for $x \in \mathbb{R}$.

(ii) The *peak power constraint* that μ is supported on $\mathsf{X}\cap[-M,M]$ for a given $M<\infty$.

We summarize these constraints on the input distribution by writing $\mu\in\mathcal{M}(\sigma_P^2,M,\mathsf{X})$, where

$$\mathcal{M}(\sigma_P^2,M,\mathsf{X}) := \Big\{\mu\in\mathcal{M} : \langle\mu,\phi\rangle\le\sigma_P^2,\ \mu\{[-M,M]\}=1\Big\}. \tag{5}$$

The capacity of a given channel subject to these constraints is denoted $C(\sigma_P^2,M,\mathsf{X})$, and may be expressed as the value of the following nonlinear program,

$$\begin{array}{rl}\textbf{max} & I(\mu)\\ \textbf{subject to} & \mu\ \in\ \mathcal{M}(\sigma_P^2,M,\mathsf{X}).\end{array} \tag{6}$$

The existence of a solution to (6) requires some conditions on the channel and its constraints. We list here the remaining assumptions imposed on the real channel in this paper.

(A1) The input alphabet X is a closed subset of $\mathbb{R}$, $\mathsf{Y}=\mathbb{C}$ or $\mathbb{R}$, and $\min(\sigma_P^2,M)<\infty$.

(A2) For each $n\ge1$,

$$\lim_{|x|\to\infty} P(|Y|<n|X=x)=0$$

(A3) The function $\log(p(\,\cdot\,|\,\cdot\,))$ is continuous on $\mathsf{X}\times\mathsf{Y}$ and, for any $y\in\mathsf{Y}$, $\log(p(y|\,\cdot\,))$ is analytic within the interior of X. Moreover, g_μ is an analytic function within the interior of X, for any $\mu\in\mathcal{M}(\sigma_P^2,M,\mathsf{X})$.

We occasionally also assume,

(A4) For any distinct pair of distributions $\mu^1,\mu^2\in\mathcal{M}(\sigma_P^2,M,\mathsf{X})$ we have $F(\mu^2\mid\mu^1)>0$, where

$$F(\mu^2\mid\mu^1) := \int\left(\frac{p(y|\mu^2)}{p(y|\mu^1)}-1\right)^2 p(y|\mu^1)\,dy. \tag{7}$$

(A5) For each finite M, the mapping $\mu\to g_\mu$ is continuous from $\mathcal{M}(\sigma_P^2,M,\mathsf{X})$ to $L_\infty[-M,M]$. That is if $\mu_n\to\mu$ weakly, with $\mu_n\in\mathcal{M}(\sigma_P^2,M,\mathsf{X})$ for all n, then

$$\lim_{n\to\infty}\sup_{|x|\le M}|g_{\mu_n}(x)-g_\mu(x)|=0. \tag{8}$$

Condition (A4) simply expresses the assumption that distinct input distributions give rise to distinct output distributions. We find in Proposition 2.6 that the mutual information $I(\mu)$ is strictly concave on $\mathcal{M}(\sigma_P^2,M,\mathsf{X})$ under this assumption. The uniform continuity (A5) holds for many channels, such as the Rayleigh channel (see (13), and the subsequent analysis of this example.)

In many examples in which $\mathsf{Y}=\mathbb{R}$ we may assume that the channel is *symmetric*. That is, $\mathsf{X}=-\mathsf{X}$, and $p(y\mid x)=p(-y\mid -x)$ for all $x,y\in\mathbb{R}$.

Proposition 2.2 *Suppose that (A1)-(A5) hold and that $M\le\infty$. Then, the mapping $I:\mathcal{M}(\sigma_P^2,M,\mathsf{X})\mapsto\mathbb{R}_+$ is continuous.*

Proof Suppose $\mu_n \to \mu$ weakly as $n \to \infty$, with $\mu_n \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ for all n. Write for each $n \geq 1$,

$$\begin{aligned} I(\mu_n) - I(\mu) &= \langle \mu_n, g_{\mu_n} \rangle - \langle \mu, g_\mu \rangle \\ &= \langle \mu_n, g_{\mu_n} - g_\mu \rangle + \langle \mu_n - \mu, g_\mu \rangle \end{aligned}$$

The second term vanishes as $n \to \infty$ by weak convergence, the continuity of g_μ (by (A3)), and the assumption that $M \leq \infty$.

The first term vanishes by (A5):

$$|\langle \mu_n, g_{\mu_n} - g_\mu \rangle| \leq \sup_{|x| \leq M} |g_{\mu_n}(x) - g_\mu(x)| \to 0, \qquad \text{as } n \to 0.$$

□

If (A5) is relaxed then we still find that I is upper semi-continuous. Consequently, there exists at least one optimal input distribution:

Proposition 2.3 *The following hold under (A1)-(A3):*

(i) *The set $\mathcal{M}(\sigma_P^2, M, \mathsf{X}) \subset \mathcal{M}$ is compact with respect to the topology of weak convergence.*

(ii) *The functional $I\colon \mathcal{M}(\sigma_P^2, M, \mathsf{X}) \to \mathbb{R}$ is concave.*

(iii) *If $M < \infty$ then I is upper semi-continuous, and an optimizer $\mu^* \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ exists.*

(iv) *If $M < \infty$, $\mathsf{Y} = \mathbb{R}$ and the channel is symmetric, then there exists an optimizer μ^* that is symmetric, in the sense that $\mu^*\{A\} = \mu^*\{-A\}$ for all $A \in \mathcal{B}(\mathsf{X})$.*

Proof Part (i) is standard (see e.g. [5]), and (ii) follows directly from Proposition 2.1.

To prove (iii), note that I is upper semi-continuous if

$$\Big\{\mu : I(\mu) \geq \gamma\Big\} \text{ is closed, for all } \gamma > 0.$$

From (4) we obtain,

$$\begin{aligned} I(\mu^\circ) \geq \gamma &\iff \langle \mu^\circ, g_\mu \rangle \geq \gamma \text{ for all } \mu \\ &\iff \mu^\circ \in \bigcap_\mu \Big\{\nu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}) : \langle \nu, g_\mu \rangle \geq \gamma\Big\}. \end{aligned} \tag{9}$$

The mapping from $\mathcal{M}(\sigma_P^2, M, \mathsf{X})$ to $\mathbb{R}$ defined by $\gamma \mapsto \langle \gamma, g_\mu \rangle$ is continuous whenever $M \leq \infty$. Consequently, the right hand side of (4) is closed, and upper semi-continuity of I follows.

The existence of an optimizer μ^* follows: For each $n \geq 1$, let $\mu_n \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ satisfy $I(\mu_n) \geq C - \frac{1}{n}$, where

$$C := \sup_{\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})} I(\mu).$$

Compactness of $\mathcal{M}(\sigma_P^2, M, \mathsf{X}) \subseteq \mathcal{M}$ implies that there exists a subsequence $\{n_k\}$ and a distribution $\mu_\infty \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ such that

$$\mu_{n_k} \to \mu_\infty, \text{ as } k \to \infty.$$

Upper semi-continuity of I then gives the upper bound,

$$C = \limsup_{k\to\infty} I(\mu_{n_k}) \le I(\mu_\infty),$$

which shows that μ_∞ is optimal.

Now we prove part (iv). If the channel is symmetric, and if μ° is any optimal input distribution, then the input distribution μ^* defined by

$$\mu^*\{A\} := \tfrac{1}{2}[\mu^\circ\{A\} + \mu^\circ\{-A\}], \qquad A \in \mathcal{B}(\mathsf{X}),$$

is also optimal, by concavity of $I(\,\cdot\,)$, and this distribution is evidently symmetric. ☐

A complex channel model is more realistic in the majority of applications. We describe next a general complex model, defined by a transition density $p_\bullet(v|u)$ on $\mathbb{C} \times \mathbb{C}$. The input is denoted U, the output V, with $U \in \mathsf{U} =$ a closed subset of $\mathbb{C}$, and $V \in \mathsf{V} = \mathbb{C}$. The input and output are related by the transition density via,

$$\mathsf{P}\{V \in dv \mid U = u\} = p_\bullet(v|u)\, dv, \qquad u, v \in \mathbb{C}.$$

The optimization problem (6) is unchanged: The average power constraint is given by $\mathsf{E}[|U|^2] \le \sigma_P^2$, and the peak-power constraint indicates that $|U| \le M$ a.s., where $|z|$ denotes the modulus of a complex number $z \in \mathbb{C}$.

We say that the complex channel model is *symmetric* if the following conditions hold:

The transition density on $\mathbb{C} \times \mathbb{C}$ satisfies,

$$p_\bullet(v|u) = p_\bullet(e^{j\alpha}v|e^{j\alpha}u), \qquad u, v \in \mathbb{C},\ \alpha \in \mathbb{R}.$$

Moreover, the constraint set U for U is *symmetric*: $\mathsf{U} = e^{j\alpha}\mathsf{U}$ for all $\alpha \in \mathbb{R}$.

In many applications it is physically reasonable to assume symmetry since phase information is lost at high bandwidths.

Under (10) we define,

(i) $X = |U|$, $\mathsf{X} = \mathsf{U} \cap \mathbb{R}_+$, and $\mathcal{M}$ denotes probability distributions on $\mathcal{B}(\mathsf{X})$;

(ii) For any $\mu \in \mathcal{M}$, we define $\mu_\bullet$ as the symmetric distribution on $\mathbb{C}$ whose magnitude has distribution μ. That is, we have the polar-coordinates representation,

$$\mu_\bullet(dx \times d\alpha) = \frac{1}{2\pi x}\mu(dx)d\alpha\,, \qquad x > 0,\ 0 \le \alpha \le 2\pi,$$

and we set $\mu(\{0\}) = \mu_\bullet(\{0\})$. This is denoted $\mu_\bullet^x$ in the special case $\mu = \delta_x$. For each $x \in \mathsf{X}$, the distribution $\mu_\bullet^x$ coincides with the uniform distribution on the circle $\{z \in \mathbb{C} : |z| = x\}$.

(iii) The transition density $p(\,\cdot\,|\,\cdot\,)$ on $\mathbb{C} \times \mathsf{X}$ is defined by

$$p(y|x) := p_\bullet(y|\mu_\bullet^x), \qquad x \in \mathsf{X},\ y \in \mathbb{C}. \tag{11}$$

(iv) $g_\mu \colon \mathbb{R}_+ \to \mathbb{R}_+$ is defined as the channel sensitivity function corresponding to the transition density p. This may be expressed,

$$g_\mu(x) = D(p_\bullet(\cdot|x) \| p_\bullet(\cdot|\mu_\bullet))\,, \qquad x \in \mathbb{R}_+\,,$$

where $\mu_\bullet$ and μ correspond as in (ii).

Proposition 2.4 justifies a reduction to the input alphabet X under (10). Under appropriate smoothness conditions on $p_\bullet$, this then provides a reduction to the special case (A1)-(A3) considered in this paper.

Proposition 2.4 *Suppose that (10) holds, and that (A1)–(A3) hold for the real channel with transition density given in (11). Then,*

(i) *For any circularly symmetric input distribution $\mu_\bullet$ for the complex channel with transition density $p_\bullet$, the mutual information may be expressed,*

$$I(\mu_\bullet) = \langle \mu, g_\mu \rangle,$$

where μ denotes the distribution of $X = |U|$.

(ii) *If $M < \infty$, then an optimal distribution $\mu_\bullet^*$ exists, and it may be taken to be symmetric. That is, for all $A \in \mathcal{B}(\mathbb{C})$ and all $\alpha \in \mathbb{R}$,*

$$\mu_\bullet^*\{A\} = \mu_\bullet^*\{e^{j\alpha}A\}.$$

Proof Part (i) follows from the observation that

$$D(p_\bullet(\cdot|x) \| p_\bullet(\cdot|\mu_\bullet)) = D(p_\bullet(\cdot|e^{j\alpha}x) \| p_\bullet(\cdot|\mu_\bullet)), \qquad x \in \mathbb{R}_+,\ \alpha \in \mathbb{R}.$$

The proof of (ii) is similar to the proof of Proposition 2.3 (iv). First, note that under the conditions of (ii) there exists an optimal distribution $\mu_\bullet^0$ by Proposition 2.3 (iii). For each $\alpha \in \mathbb{R}$, define a new distribution $\mu_\bullet^\alpha$ on $\mathcal{B}(\mathbb{C})$ by

$$\mu_\bullet^\alpha\{A\} := \mu_\bullet^0\{e^{j\alpha}A\}, \qquad A \in \mathcal{B}(\mathbb{C})\,,$$

and set $\mu_\bullet^* = \frac{1}{2\pi}\int_0^{2\pi} \mu_\bullet^\alpha\, d\alpha$. Under (10) we must have $I(\mu_\bullet^0) = I(\mu_\bullet^\alpha)$ for each $\alpha \in \mathbb{R}$, and hence $\mu_\bullet^*$ must also be optimal by concavity of I. ☐

Proposition 2.4 justifies the consideration of a channel with real input alphabet, in which $\mathsf{X} = \mathbb{R}_+$ and $\mathsf{Y} = \mathbb{C}$. Throughout the remainder of the paper, in all of our analysis we restrict to a channel satisfying (A1)–(A3) with real input-alphabet.

***Example*: The Rician channel**

This is the general complex fading channel, in which the input and output are related by,

$$V = (A + a)U + N \tag{12}$$

where U and V are the complex-valued channel input and output, $a \geq 0$, and A and N are independent complex Gaussian random variables, $A \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_A^2)$ and $N \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_N^2)$. Throughout the paper we assume that N and A are circularly symmetric. Consequently, V has a circularly symmetric distribution whenever the distribution of U is circularly symmetric.

The Rician channel reduces to the complex AWGN channel when $\sigma_A^2 = 0$. ☐

On setting $a = 0$ we obtain another important special case:

***Example*: The Rayleigh channel**

The model (12) with $a = 0$ is known as the Rayleigh channel. Under our standing assumption that N, A have circularly symmetric distributions, it follows that the output distribution is symmetric for *any* input distribution (not necessarily symmetric.)

Based on this property, the model may be normalized as follows, as in [1]: Setting $X = |U|\sigma_A/\sigma_N$ and $Y = |V|^2/\sigma_N^2$, we obtain a real channel model with transition density

$$p(y|x) = \frac{1}{1+x^2} \exp\Big(-\frac{1}{1+x^2}\, y\Big), \qquad x, y \in \mathbb{R}_+. \tag{13}$$

The sensitivity function g_μ is easily computed numerically for a given $\mu \in \mathcal{M}$ based on (13). For the general Rician model, computation of g_μ appears to be less straightforward: This requires computation of $g_{\mu_\bullet}$, which involves integration over the complex plane. ☐

Proposition 2.5 *Proposition (A1)-(A5) hold for the Rayleigh channel.*

Proof Proposition (A1)-(A3) are easily established.

From [1, App. I, Lem. 2], we have for each pair $\mu^1, \mu^2 \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$, if the densities coincide so that

$$p(y|\mu^2) = p(y|\mu^1) \text{ for all } y \in [0, M],$$

then $\mu^2 = \mu^1$. This shows that (A4) holds for the Rayleigh channel.

Next we establish (A5). Suppose that $\{\mu_n\} \subset \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ converge weakly to some $\mu_\infty \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$. For each $n \geq 1$ we express the channel discrimination function as follows,

$$g_{\mu_n}(x) = \int \log(p(y|x))p(y|x)dy + \int \log(p(y|\mu_n))p(y|x)dy. \tag{14}$$

The first term on the right hand side does not depend on μ_n, so we only need to consider the second term. For each $y \in \mathbb{R}$ the density $p(y|x)$ is a bounded and continuous function of $x \in [0, M]$. Consequently, since$\mu_n \to \mu_\infty$ weakly,

$$\lim_{n\to\infty} \log(p(y|\mu_n)) \to \log(p(y|\mu_\infty)), \qquad \text{as } n \to \infty\,, \quad y \in \mathbb{R}. \tag{15}$$

Moreover, applying (13), we obtain the uniform bound,

$$|\log(p(y|\mu))| \le y + k_0, \qquad y \in \mathbb{R},\ \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}), \tag{16}$$

where the constant k_0 only depends on M. Hence, by (14), (15) and the Dominated Convergence Theorem,

$$g_{\mu_n}(x) \to g_{\mu_\infty}(x) \text{ for each } x. \tag{17}$$

To complete the proof that (A5) holds we now strengthen the pointwise convergence in (17) to uniform convergence on $[0, M]$. Similar to (16), it may be shown that for any $M < \infty$, there exists $k_1 < \infty$ such that for all $x \in [0, M]$ and all $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$,

$$\Big|\frac{d}{dx} g_\mu(x)\Big| \le k_1.$$

It follows that $\{g_{\mu_n} : n \ge 1\}$ is equicontinuous on $[0, M]$. Ascoli's Theorem then implies the desired uniform convergence. $\square$

Since $\mathcal{M}$ is a convex set, and I a concave functional on $\mathcal{M}$, computation of capacity may be viewed as a convex optimization problem. We turn to structural properties of its dual next to obtain characterizations of optimal distributions, and sensitivity formulae.

2.2 Kuhn-Tucker conditions

Proposition 2.1 implies that g_μ is the gradient of I at μ [19]. The next result expresses this observation in terms of directional derivatives, and establishes a formula for the second derivative that may be viewed as a form of Fisher information. Let μ° be a fixed element of $\mathcal{M}(\sigma_P^2, M, \mathsf{X})$ and θ a real number in $[0, 1]$. For any $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$, define $\tilde{\mu} = \mu - \mu^\circ$ and $\mu_\theta := (1-\theta)\mu^\circ + \theta\mu = \mu^\circ + \theta\tilde{\mu}$. The sensitivity of mutual information along the direction from μ° to μ is quantified in the following proposition.

Proposition 2.6 *The first and second order sensitivities of mutual information with respect to the input distribution are given by, respectively,*

$$\frac{d}{d\theta} I(\mu_\theta)\Big|_{\theta=0} = \langle \mu - \mu^\circ, g_{\mu^\circ} \rangle, \tag{18}$$

$$\frac{d^2}{d\theta^2} I(\mu_\theta)\Big|_{\theta=0} = -F(\mu \mid \mu^\circ) \tag{19}$$

where $F(\mu \mid \mu^\circ)$ is defined in (7). This is equal to the Fisher information on $p(y|\mu_\theta)$, at $\theta = 0$.

PROOF For any θ we have from the definition of mutual information,

$$I(\mu_\theta) = -\iint p(y|x) \ln\left(\frac{p(y|\mu_\theta)}{p(y|x)}\right) dy\ \mu_\theta(dx)$$

and differentiating with respect to θ gives

$$\begin{aligned}
\frac{d}{d\theta} I(\mu_\theta) &= -\iint p(y|x) \ln\left(\frac{p(y|\mu_\theta)}{p(y|x)}\right) dy\, \tilde{\mu}(dx) \\
&\quad - \iint p(y|x) \left(\frac{p(y|x)}{p(y|\mu_\theta)}\right) \left(\frac{p(y|\tilde{\mu})}{p(y|x)}\right) dy\, \mu_\theta(dx) \\
&= \iint p(y|x) \ln\left(\frac{p(y|x)}{p(y|\mu_\theta)}\right) dy\, \tilde{\mu}(dx). \qquad (20)
\end{aligned}$$

In (20) we have abused notation slightly, writing $p(y|\tilde{\mu}) := p(y|\mu) - p(y|\mu^\circ)$.

The following identities follow directly from (20):

$$\begin{aligned}
\frac{d}{d\theta} I(\mu_\theta) &= \int g_{\mu_\theta}(x) \tilde{\mu}(dx), \\
\frac{d^2}{d\theta^2} I(\mu_\theta) &= -\iint p(y|x) \frac{p(y|\tilde{\mu})}{p(y|\mu_\theta)}\, dy\, \tilde{\mu}(dx).
\end{aligned}$$

Evaluating at $\theta = 0$ then gives (18) and (19). □

The first order sensitivity formula given in (18) is similar to the expression for mutual information given in (4). These expressions form the basis for the new capacity computation algorithms proposed in Section 4.

Now we turn to structural properties of the convex program (6) and its convex dual to obtain characterizations of optimal distributions.

The *dual functional* $\Psi\colon \mathbb{R}_+ \to \mathbb{R}_+$ is defined by

$$\Psi(r) = \sup_{\mu \in \mathcal{M}_0} [I(\mu) - r\langle \mu, \phi \rangle], \qquad r \geq 0, \qquad (21)$$

where $\mathcal{M}_0 = \mathcal{M}(M^2, M, \mathsf{X}) = \mathcal{M}(\infty, M, \mathsf{X})$ denotes the constraint set without an average power constraint. The dual functional is a convex, decreasing function of r, as illustrated in Figure 2.2. Note that we do not exclude $M = \infty$. In this case, $\mathcal{M}_0 = \mathcal{M}$, which denotes the set of probability distributions on X.

The proof of the following result is identical to the proof of Proposition 2.3 (iii).

Proposition 2.7 *Suppose that (A1)-(A3) hold, and that $M < \infty$. Then for each $r > 0$ there exists an optimizer $\mu_r^* \in \mathcal{M}$ satisfying $\Psi(r) = [I(\mu_r^*) - r\langle \mu_r^*, \phi \rangle]$.* □

The parameter r provides a convenient parameterization of the optimization problem (6). This is made clear in Theorem 2.8 and its corollary, Proposition 2.9. Similar techniques are used in the theoretical development of the random coding error exponent [15, 7, 8].

Theorem 2.8 *Suppose an optimizing distribution μ_r^* exists for (21) for each $r > 0$. The following then hold:*

(i) $\Psi(r) = \min_{\mu \in \mathcal{M}_0} \|[g_\mu - r\phi]_+\|_\infty$.

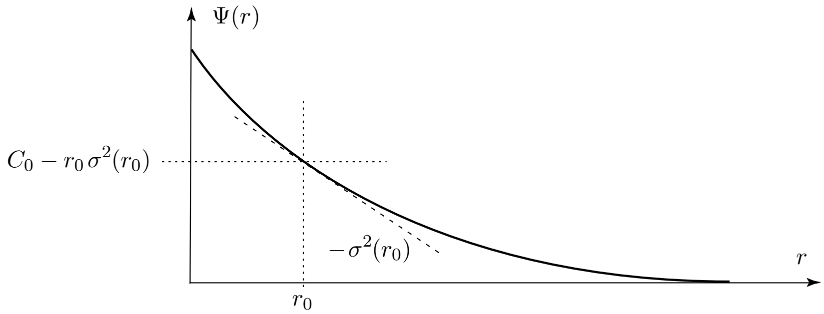

Figure 1: The dual functional is convex and decreasing. For a given $r_0 > 0$, the slope determines an average power constraint $\sigma^2(r_0)$, and the corresponding capacity $C_0 := C(\sigma^2(r_0), M, \mathsf{X})$ may be determined as shown in the figure.

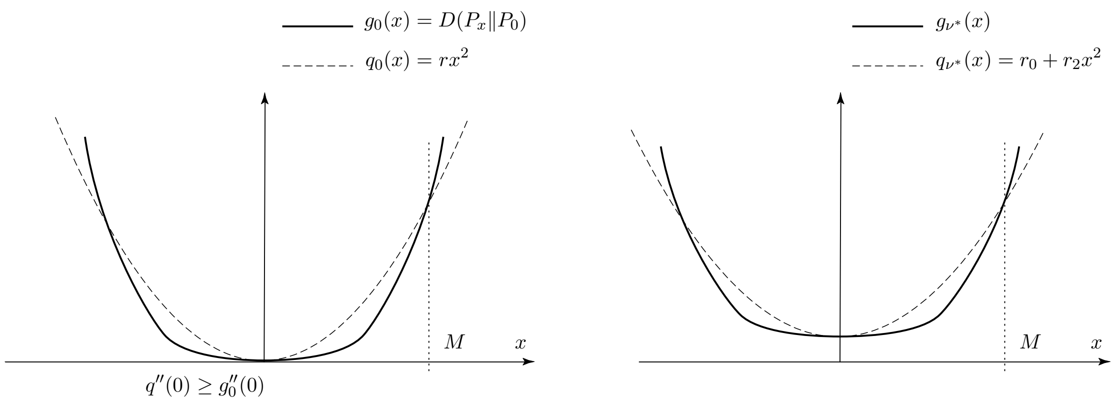

Figure 2: The left hand side shows alignment of g_0 with a pure quadratic. When δ_0 is perturbed to form ν^*, the functions g_0 and q_0 are perturbed as shown at right.

(ii) *Let* $\sigma^2(r) := -\frac{d}{dr}\Psi(r)$. *The distribution* μ_r^* *is optimal under the corresponding average power constraint:*

$$I(\mu_r^*) = C(\sigma^2(r), M, \mathsf{X}).$$

Moreover, we have

$$I(\mu_r^*) = \Psi(r) + r\sigma^2(r)$$

(iii) *The capacity* $C(\,\cdot\,, M, \mathsf{X})$ *is concave in its first variable, with*

$$\frac{d}{d\sigma_P^2} C(\sigma_P^2, M, \mathsf{X}) = r, \qquad \textit{when } \sigma_P^2 = \sigma^2(r).$$

Proof To prove part (i), we first apply Proposition 2.1 as follows: For fixed $r > 0$,

$$\begin{aligned}\Psi(r) &= \sup_{\mu\in\mathcal{M}_0} [I(\mu) - r\langle\mu,\phi\rangle] \\ &= \sup_{\mu\in\mathcal{M}_0} \inf_{\mu'\in\mathcal{M}_0} \langle\mu, g_{\mu'} - r\phi\rangle \\ &\le \inf_{\mu'\in\mathcal{M}_0} \sup_{\mu\in\mathcal{M}_0} \langle\mu, g_{\mu'} - r\phi\rangle \\ &= \inf_{\mu'\in\mathcal{M}_0} \sup_{x\in\mathsf{X},\,|x|\le M} (g_{\mu'}(x) - rx^2) \\ &= \inf_{\mu\in\mathcal{M}_0} \|[g_\mu - r\phi]_+\|_\infty .\end{aligned}$$

Conversely, suppose μ^*_r is optimal, fix $|x| \le M$, and for $0 \le \theta \le 1$, let $\mu_\theta = (1-\theta)\mu^*_r + \theta\delta_x$. By optimality,

$$\frac{d}{d\theta}\left(I(\mu_\theta) - r\langle\mu_\theta,\phi\rangle\right)\Big|_{\theta=0} \le 0,$$

which gives, on applying Proposition 2.6,

$$g_{\mu^*_r}(x) - rx^2 \le I(\mu^*_r) - r\sigma^2_r = \Psi(r).$$

This completes the proof of (i) since x is arbitrary.

Part (ii) follows from [19, Sec. 8.3, Thm. 1], and (iii) follows from [19, Sec. 8.5, Thm. 1]. □

Proposition 2.8 leads to the following version of the Kuhn-Tucker alignment conditions. A related result is presented as [1, Theorem 4]. See also [15, 8].

Proposition 2.9 *The following hold under (A1)–(A3):*

(i) *Suppose that an optimizing distribution μ^*_r exists. Then, setting q equal to the quadratic function $q = \Psi(r) + r\phi$, we have,*

$$\begin{aligned} g_{\mu^*_r}(x) &\le q(x), \qquad x \in \mathsf{X}; \\ g_{\mu^*_r}(x) &= q(x), \qquad a.e.\ [\mu^*_r] \end{aligned} \tag{22}$$

(ii) *Suppose that $\mu^\circ \in \mathcal{M}_0$, and that constants $\varepsilon > 0$, $\sigma^2_P > 0$ together with a quadratic function $q = r_0 + r_2\phi$ exist, satisfying*

(a) $\langle\mu^\circ,\phi\rangle = \sigma^2_P$ (b) $g_{\mu^\circ} \le q$ on $\mathsf{X} \cap [-M, M]$ (c) $\langle\mu^\circ, q - g_{\mu^\circ}\rangle \le \varepsilon$.

Then,

$$I(\mu^\circ) \ge C(\sigma^2_P, M, \mathsf{X}) - \varepsilon.$$

(iii) *Suppose that $\mu^\circ \in \mathcal{M}_0$, and that $\varepsilon > 0$, $r_0 > 0$ exist, satisfying*

$$g_{\mu^\circ} \le r_0, \quad \text{and} \qquad \langle\mu^\circ, g_{\mu^\circ}\rangle \ge r_0 - \varepsilon.$$

Then,

$$I(\mu^\circ) \ge C(M^2, M, \mathsf{X}) - \varepsilon.$$

Proof   From Theorem 2.8 (i) we have

$$\Psi(r) \geq g_{\mu_r^*}(x) - rx^2\,, \qquad x \in \mathsf{X},$$

and by optimality $\Psi(r) = \langle \mu_r^*, g_{\mu_r^*} - r\phi \rangle$. Part (i) easily follows.

To prove part (ii), define $\mu_\theta := (1-\theta)\mu^\circ + \theta\mu$ for $\theta \in [0,1]$. From the assumption that $\langle \mu^\circ, q \rangle = \langle \mu, q \rangle$ we have,

$$\begin{aligned} \frac{dI(\mu_\theta)}{d\theta}\Big|_{\theta=0} &= \langle \mu - \mu^\circ, g_{\mu^\circ} \rangle \\ &= \langle \mu - \mu^\circ, g_{\mu^\circ} - q \rangle \\ &= \langle \mu, g_{\mu^\circ} - q \rangle - \langle \mu^\circ, g_{\mu^\circ} - q \rangle \leq \epsilon. \end{aligned}$$

Because the mutual information is a concave function over μ, it follows that

$$I(\mu) \leq I(\mu^\circ) + \frac{dI}{d\theta}(\mu_\theta)\Big|_{\theta=0} \leq I(\mu^\circ) + \epsilon.$$

Since $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ is arbitrary, this establishes (ii).

Part (iii) follows similarly:

$$\begin{aligned} \frac{dI(\mu_\theta)}{d\theta}\Big|_{\theta=0} &= \langle \mu - \mu^\circ, g_{\mu^\circ} \rangle \\ &= \langle \mu - \mu^\circ, g_{\mu^\circ} - r_0 \rangle \\ &= \langle \mu, g_{\mu^\circ} - r_0 \rangle - \langle \mu^\circ, g_{\mu^\circ} - r_0 \rangle \leq \epsilon. \end{aligned}$$

Consequently, the proof follows from concavity of $I(\,\cdot\,)$, as in (ii). ☐

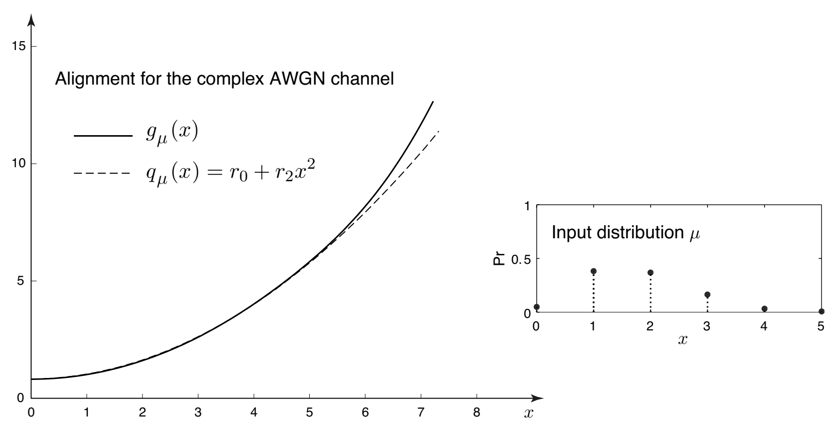

Figure 3: The sensitivity function g_μ for the complex AWGN channel for the distribution μ shown in Figure 8, step 35.

***Example*: Alignment for the complex AWGN channel**

An illustration of Proposition 2.9 is provided by Figure 3. The channel considered is the complex AWGN channel, whose statistics are shown in (32). The distribution μ on the magnitude of the input is obtained from Figure 8, step 35. It is discrete, with five points of support.

The sensitivity function g_μ is in nearly perfect alignment with a quadratic on the interval $[0, 5]$. For $x > 5$ we see that g_μ is greater than this quadratic, from which we conclude that μ is not optimal for $\mathsf{X} = \mathbb{R}_+$, but it is very nearly optimal when X is equal to the interval $[0, 5]$. $\square$

3 Why Are Optimal Distributions Discrete?

As surveyed in the introduction, it is known that the capacity-achieving distribution is discrete in many channel models. In fact, the AWGN channel under an average power constraint, with $M = \infty$ is the only example we know of in which the optimal input distribution is absolutely continuous with respect to Lebesgue measure. Here we consider the general channel model satisfying (A1)-(A3) and provide a series of results and examples to explain why the optimizer $\mu^* \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ for (6) is typically discrete.

Throughout this section we take $\mathsf{X} = \mathbb{R}_+$.

3.1 Alignment

For a channel subject to only the peak power constraint, we have

Proposition 3.1 *Suppose that (A1)–(A3) hold with $\mathsf{X} = \mathbb{R}_+$, and with $\sigma_P^2 = \infty$, $M < \infty$. Then, there exists an optimal input distribution μ^* which is discrete, with a finite number of mass points.*

Proof Existence of μ^* follows from Proposition 2.3 (iii).

We show next that the channel sensitivity function g_μ is unbounded for any input distribution μ. In particular, this holds for $\mu = \mu^*$. Since g_{μ^*} is assumed to be an analytic function on X, and it is not a constant function, it then follows that $g_{\mu^*}(x) = I(\mu^*)$ for only a finite number of $x \in \mathsf{X} \cap [-M, M]$. We conclude from Proposition 2.9 that the optimal input distribution is discrete, with a finite number of mass points.

We now show that g_μ is unbounded. For any fixed $n \geq 1$, define $b_n > 0$ by

$$b_n^{-1} := \int_{|y| \geq n} p(y|\mu)\, dy.$$

We note that that $b_n p(y|\mu) \mathbf{1}(|y| \geq n)$ defines a probability density on Y.

We define $h : \mathbb{R}_+ \to \mathbb{R}$ by

$$h(z) = z \log(z), \qquad z \in \mathbb{R}_+.$$

This is a convex function for $z \in \mathbb{R}^+$, and is bounded from below by $-e^{-1}$.

The channel sensitivity function may be bounded from below as follows, for each $n \geq 1$:

$$\begin{aligned} g_\mu(x) &= \int_{|y|\leq n} \log\left(\frac{p(y|x)}{p(y|\mu)}\right)\left(\frac{p(y|x)}{p(y|\mu)}p(y|\mu)\right) dy + \int_{|y|>n} \log\left(\frac{p(y|x)}{p(y|\mu)}\right)\frac{p(y|x)}{p(y|\mu)}p(y|\mu)\, dy \\ &\geq \inf_{z\in\mathbb{R}} h(z) + b_n^{-1}\int_{|y|>n} h\left(\frac{p(y|x)}{p(y|\mu)}\right)[b_n p(y|\mu)]\, dy \\ &\geq -e^{-1} + b_n^{-1} h\left(b_n \mathsf{P}_x(|Y| > n)\right), \end{aligned}$$

where the last inequality follows from Jensen's inequality.

From this bound and (A3) we conclude that for any $n \geq 1$,

$$\liminf_{|x|\to\infty} g_\mu(x) \geq -e^{-1} + b_n^{-1} h(b_n) = -e^{-1} + \log(b_n).$$

Since $b_n \to \infty$, as $n \to \infty$, it follows that g_μ is unbounded. $\square$

We have the following corollary for the symmetric complex channel. The assumptions of Corollary 3.1 hold for many models, including the AWGN, Rayleigh, and Rician channels. The proof follows directly from Proposition 2.4 and Proposition 3.1.

Corollary 3.1 *Consider a complex channel model satisfying (10). Suppose that (A1)–(A3) hold for the transition density given in (11). Assume moreover that $\sigma_P^2 = \infty$, and $M < \infty$. Then, there exists an optimal input distribution $\mu_\bullet^*$ on $\mathcal{B}(\mathbb{C})$ whose phase is uniformly distributed and independent of magnitude. Its magnitude μ^* is discrete, with a finite number of mass points in $[0, M]$.* $\square$

3.2 Optimal binary distributions

We now consider the case of vanishing SNR. Recall that we restrict to $\mathsf{X} = \mathbb{R}_+$.

We say that $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ is *binary* if it has two points of support in $[0, M]$. It is known that a binary distribution is *approximately* optimal in certain limiting regimes. For example, it is shown in [14] that the restriction to binary inputs is essentially optimal for a discrete-time, point-to-point channel with an input alphabet $\{0, 1, ..., K\}$, in the broadband limit as SNR goes to zero. Similarly, a binary input is approximately optimal when 'bits are sufficiently inexpensive' [31, Theorem 3].

The conclusions of both [31] and [14] may be interpreted via properties of extreme points in an associated infinite-dimensional linear program. To see this, fix any symmetric $\mu_0 \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$, and conclude by Proposition 2.1 that $C(\sigma_P^2, M, \mathsf{X})$ is bounded from above by the solution to the linear program,

$$\begin{array}{rl} \textbf{max} & \langle \mu, g_{\mu_0} \rangle \\ \textbf{subject to} & \langle \mu, \phi \rangle \leq \sigma_P^2 \\ & \langle \mu, \boldsymbol{1} \rangle \leq 1 \end{array} \tag{23}$$

where the variable μ in (23) is a positive measure on $[0, M]$. An optimal solution to this linear program can be taken with at most *two points of support* since this corresponds to a basic feasible solution. That is, the optimizer of (23) is binary.

The linear program (23) may be interpreted as a relaxation of the original convex program (6). We show in Theorem 3.3 that, for an appropriate choice of μ_0, this relaxation is a minor perturbation when $\sigma_P^2 \sim 0$. Based on this approximation, we conclude in Theorem 3.3 that for $\sigma_P^2 > 0$ sufficiently small, the optimal input distribution for (6) is in fact binary.

The next question is, *how should we choose the distribution* μ_0? Since optimizer of (23) has average power $\sigma_P^2 \sim 0$, it is reasonable to take $\mu_0 = \delta_0$, the point mass at the origin. The function g_{δ_0} is then precisely the channel discrimination function g_0 defined in (3). We let ν^* denote the (symmetric) binary optimizer of the linear program (23) with $g_{\mu_0} = g_0$.

Under (A1)–(A5) we may conclude that $g_\mu \approx g_0$ for small values of σ_P^2 whenever $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ with $\langle \mu, \phi \rangle \leq \sigma_P^2$. It follows that $\langle \mu, g_0 \rangle \approx \langle \mu, g_\mu \rangle$ for such μ, which strongly suggests that ν^* is nearly optimal. This approximation is made precise in Theorem 3.3. We first consider an example.

***Example*: Sensitivity for the Rayleigh channel**

For the Rayleigh channel, normalized according to (13), the channel discrimination function is given by

$$g_0(x) = x^2 - \log(1 + x^2), \qquad x \in \mathbb{R}_+. \tag{24}$$

From this expression it follows that $g_0'(0) = g_0''(0) = 0$. Consequently, applying Proposition 2.6, we see that first and second order sensitivity are extremely low when μ is supported on a small interval containing the origin, and in this case $I(\mu)$ is also extremely small.

This observation suggests that the optimal input distribution may take on very large values, albeit with small probability. □

In addition to a peak power constraint, the following restriction is imposed in Theorem 3.3:

$$\frac{d \log(g_0(x))}{d \log(x)} > 2, \qquad 0 < x \leq M. \tag{25}$$

This implies that the growth rate of g_0 is faster than any quadratic, in the sense that

$$\frac{d}{dx} \log(g_0(x)) > \frac{d}{dx} \log(rx^2), \qquad r > 0\,,\, 0 < x < M\,.$$

The conclusions of Lemma 3.2 are the only structural properties required in the proof of Theorem 3.3.

Lemma 3.2 *Suppose that (25) holds. Then, setting $r(0) = g_0(M)/M^2$, and $q_0 := r(0)\phi$, we obtain the following conclusions:*

(a) $g_0(0) = q_0(0)$ *and* $g_0(M) = q_0(M)$;

(b) $g_0(x) < q_0(x)$ *for* $x \in (0, M)$;

(c) $g_0'(M) > q_0'(M)$;

(d) $g_0'(0) = q_0'(0) = 0$, *and* $g_0''(0) < q_0''(0) = 2g_0(M)/M^2$.

Proof All of these conclusions are obvious, except for the bound $g_0''(0) < 2g_0(M)/M^2$, which is demonstrated here.

Let $H(x) = \log(x^{-2}g_0(x))$, $0 < x \le M$, and define $H(0) = \lim_{x\downarrow 0} H(x)$. An application of L'Hopital's rule gives $H(0) = \log(g_0''(0)/2)$.

Under (25) we have $H'(s) > 0$ on $(0, M]$ and we can conclude that

$$\log(g_0(M)/M^2) := H(M) > H(x)\,, \qquad 0 < x < M\,.$$

Letting $x \downarrow 0$ then gives $\log(g_0(M)/M^2) > H(0) = \log(g_0''(0)/2)$, which is the desired conclusion. ☐

The following result provides conditions under which the optimal input distribution is binary, as well as sensitivity bounds with respect to σ_P^2 for low SNR. Theorem 3.3 also shows that the discrimination function gives a strict upper bound on capacity for a peak power constrained channel satisfying (25).

Theorem 3.3 *Suppose that Conditions (A1)–(A4) hold, together with the bound (25). Assume moreover that* $\mathsf{X} = \mathbb{R}_+$. *Then, there exists* $\bar\sigma_P^2 > 0$ *such that the following hold for* $0 < \sigma_P^2 \le \bar\sigma_P^2$:

(i) ν^* *is equal to the unique binary distribution on* $\{0, M\}$ *satisfying* $\langle \nu^*, \phi\rangle = \sigma_P^2$.

(ii) *This binary distribution* ν^* *is optimal. That is,* $I(\nu^*) = C(\sigma_P^2, M, \mathsf{X})$.

(iii) *The sensitivity with respect to the average power constraint satisfies,*

$$\lim_{\sigma_P^2\downarrow 0}\Big(\frac{d}{d\sigma_P^2} C(\sigma_P^2, M, \mathsf{X})\Big) = g_0(M)/M^2 > g_0''(0).$$

Consequently, for all σ_P^2,

$$\begin{aligned} C(\sigma_P^2, M, \mathsf{X}) \quad &\le \quad \max\Big\{\frac{g_0(M)}{M^2}\sigma_P^2,\, g_0(M)\Big\} \\ &= \quad g_0(M)\max\Big\{\frac{\sigma_P^2}{M^2}, 1\Big\} \end{aligned}$$

Proof For $\theta \in [0,1]$ define $\nu_\theta \in \mathcal{M}_0$ by,

$$\nu_\theta := (1-\theta)\delta_0 + \theta\delta_M\,, \qquad g_\theta := g_{\nu_\theta}.$$

For each fixed $x \in \mathbb{R}$, $g_\theta(x)$ is a non-negative, convex function of θ on $[0,1]$. For each θ we let q_θ denote the unique quadratic of the form $q_\theta = r_0(\theta) + r_2(\theta)\phi$ satisfying $q_\theta(0) = g_\theta(0)$ and $q_\theta(M) = g_\theta(M)$. Note that $q_\theta \to q_0$, $g_\theta \to g_0$ as $\theta \downarrow 0$.

Under properties (a)–(d) above it easily follows that there exists $\bar\theta > 0$ such that the following analogous conditions hold for $\theta \in (0, \bar\theta]$:

(a) $g_\theta(0) = q_\theta(0)$ and $g_\theta(M) = q_\theta(M)$;

(b) $g_\theta(x) < q_\theta(x)$ for $x \in (0, M)$;

(c) $g'_\theta(M) > q'_\theta(M)$;

(d) $g'_\theta(0) = q'_\theta(0) = 0$, and $g''_\theta(0) < q''_\theta(0)$.

In particular, the alignment conditions hold, so by Proposition 2.9 the distribution ν_θ optimizes $\Psi(r_2(\theta))$. This completes the proof of (i) and (ii), with $\bar\sigma_P^2 := \langle \nu_{\bar\theta}, \phi\rangle$.

Part (iii) follows from Lemma 3.2 and Theorem 2.8 which in particular implies concavity of $C(\cdot, M, \mathsf{X})$, and the observation that $r_2(\theta) \to r(0) = g_0(M)/M^2$ as $\theta \downarrow 0$. ☐

These results also imply conclusions for the channel optimization problem without peak power constraint. The following result is an easy consequence of Theorem 3.3.

Corollary 3.2 *Suppose that (25) holds, and consider the parameterized family of optimization problems, parameterized by $a = \sigma_P^2 > 0$. We let μ_a^* denote the optimizing distribution on X and set $M(a) = \sup\bigl(M > 0 : \mu_a^*\{[M,\infty)\} > 0\bigr)$. Then $M(a) \to \infty$, as $a \to 0$.*

☐

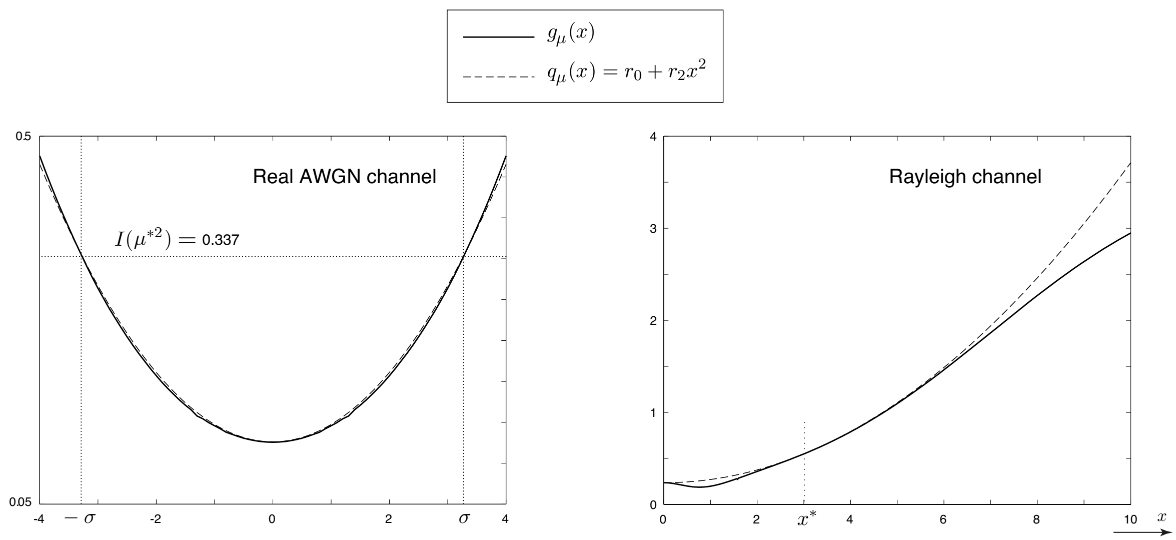

Figure 4: The sensitivity function g_μ. The illustration at left illustrates the alignment condition for the real AWGN channel, with $\mu = \frac{1}{2}(\delta_\sigma + \delta_{-\sigma})$. The illustration at right shows the alignment condition for the complex Rayleigh channel, with input distribution symmetric, with magnitude supported on $\{0, 5\}$. The channel statistics are shown in (32).

We conclude with examples illustrating the assumptions and conclusions of Theorem 3.3.

***Example*: Simple input distributions for the real AWGN channel**

To place the real AWGN channel within the context of Theorem 3.3 we first perform a reduction to the case $\mathsf{X} = \mathbb{R}_+$, just as was done in our consideration of complex channel models. In this case the input X is interpreted as the magnitude of the channel input.

However, in this model the bound (25) is violated, since the left hand side of (25) is identically 2. Nevertheless, we show here that the optimal input distribution has only one point of support (it is a *degenerate* binary distribution) for certain parameter values.

Take $\sigma_N^2 = 10$, $\sigma_P^2 = 10$, and set $\mu^{*2} := \delta_{\sigma_P}$. This is the distribution on the magnitude of X: The associated input distribution for the channel will use ± 10 with probability $\frac{1}{2}$. Figure 4 shows the function $g_{\mu^{*2}}$ and a quadratic function $q_{\mu^{*2}}$ satisfying $g_{\mu^{*2}} \leq q_{\mu^{*2}}$ on $[-\sigma_P, \sigma_P]$ with $g_{\mu^{*2}}(\sigma_P) = q_{\mu^{*2}}(\sigma_P)$. It follows that the alignment condition holds for the convex program with peak power constraint given by $M = \sigma_P$. For $M > \sigma$, the alignment condition is violated on $[-\sigma_P, \sigma_P]^c$. We conclude that μ^{*2} is the optimal distribution using the constraint set $\mathcal{M}(\sigma_P^2, M, \mathbb{R})$ with $M = \sigma_P$, but μ^{*2} is *not optimal* when $M > \sigma_P$.

The mutual information using the binary input distribution is approximately $I(\mu^{*2}) \approx 0.337$, while the capacity under an average power constraint alone is given by

$$C(10, \infty, \mathbb{R}) = \tfrac{1}{2} \ln(1 + \sigma_P^2/\sigma_N^2) \approx 0.347.$$

In conclusion, we find that restricting the input to be binary results in a capacity loss of approximately three percent. Similar conclusions are discussed in [28]. $\square$

***Example*: Binary distributions for the Rician channel**

Consider the Rician channel (12), normalized with $\sigma_N^2 = 1$. The channel discrimination function may be explicitly computed,

$$g_0(x) = (a^2 + \sigma_A^2)x^2 - \log(1 + \sigma_A^2 x^2)\,, \qquad x \in \mathbb{R}_+,$$

and hence,

$$\frac{d\log(g_0(x))}{d\log(x)} = \frac{x g_0'(x)}{g_0(x)} = 2\Big[\frac{x^2[(a^2 + \sigma_A^2)(1 + \sigma_A^2 x^2) - \sigma_A^2]}{(1 + \sigma_A^2 x^2)[(a^2 + \sigma_A^2)x^2 - \log(1 + \sigma_A^2 x^2)]}\Big].$$

When $\sigma_A^2 = 0$ we obtain the complex AWGN channel. In this case the bound (25) is violated since $d\log(g_0(x))/d\log(x) \equiv 2$.

The right hand side is strictly greater than 2, for all $0 < x < \infty$, whenever $\sigma_A^2 > 0$. Consequently, in this case the bound (25) holds, and the conclusions of Theorem 3.3 also hold for any $M < \infty$.

Figure 4 shows results from one numerical experiment for the Rayleigh channel model in which a binary distribution is optimal without a peak power constraint. The channel statistics are shown in (32), resulting in an SNR equal to 4 (i.e. 6 dB). $\square$

4 Algorithms

We now introduce new classes of algorithms to estimate capacity and construct efficient, discrete input distributions. All of these algorithms are based upon approximations of the convex program (6) with an appropriate linear program.

4.1 Cutting plane algorithm

We have already seen in Section 3.2 that a relaxation of the expression for I given in Proposition 2.1 may provide insight into the structure of optimal distributions, and even computational methods. Here we describe a general computational algorithm based on a sequence of increasingly tight relaxations of (4).

The algorithm introduced here is a special case of the cutting plane algorithm first proposed in [18]. Modern treatments of the general cutting plane algorithm may be found in [9, 3, 16].

Cutting plane algorithm

The algorithm is initialized with an arbitrary distribution $\mu_0 \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$, and inductively constructs a sequence of distributions as follows. At the nth stage of the algorithm, we are given n distributions $\{\mu_0, \mu_2, \dots, \mu_{n-1}\} \subset \mathcal{M}(\sigma_P^2, M, \mathsf{X})$. We then define,

(i) The piecewise linear approximation,

$$I_n(\mu) := \min_{0 \le i \le n-1} \langle \mu, g_{\mu_i} \rangle, \qquad \mu \in \mathcal{M}\,. \tag{26}$$

(ii) The next distribution,

$$\mu_n = \arg\max\{I_n(\mu) : \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})\}. \tag{27}$$

From Proposition 2.1 it is evident that $I_n(\mu) \ge I(\mu)$ for all $\mu \in \mathcal{M}$.

The celebrated Blahut-Arimoto algorithm is also obtained as a relaxation of the convex optimization problem (6) (see [6].) This algorithm requires a computation of $C(\sigma^2(r), M, \mathsf{X})$ for a range of $r > 0$ in order to find $C(\sigma_P^2, M, \mathsf{X})$. In numerical experiments using simple finite channel models we have found that the cutting plane algorithm converges slightly faster than the Blahut-Arimoto algorithm.

The optimization problem (27) is equivalently expressed as the solution to the linear program,

$$\begin{array}{rrcl} \textbf{max} & c & & \\ \textbf{subject to} & \langle \mu, g_{\mu_i} \rangle & \ge & c, \\ & 0 & \le & i \le n-1, \\ & \mu & \in & \mathcal{M}(\sigma_P^2, M, \mathsf{X}). \end{array} \tag{28}$$

We note that the linear program (28) is finite-dimensional only when the cardinality of X is finite.

We let (c_n, μ_n) denote the optimizer of (28). The algorithm is convergent under a peak power constraint:

Theorem 4.1 *Suppose that (A1)–(A5) hold and $M < \infty$. Then, the cutting plane algorithm generates a sequence of distributions $\{\mu_n : n \ge 1\} \subset \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ such that*

(i) $\mu_n \to \mu^*$ *weakly, as* $n \to \infty$*;*

(ii) $I(\mu_n) \to I(\mu^*) = C(\sigma_P^2, M, \mathsf{X})$*;*

(iii) $c_1 \geq c_2 \geq c_3 ... \to I(\mu^*)$*;*

(iv) μ_n *can be chosen so that it has at most* $n+1$ *points of support for each* $n \geq 1$.

PROOF Because $\mathcal{M}$ is a compact set, there exists a subsequence $\{\mu_{n_k}\}$ that converges weakly to some distribution $\mu_\infty \in \mathcal{M}$. We show here that μ_∞ is the optimal distribution. From the definitions (26) and (27), we obtain for all k, all $i < k$, and all μ

$$I(\mu_{n_i}) + \langle \mu_{n_k} - \mu_{n_i},\ g_{\mu_{n_i}} \rangle \geq I_{n_k}(\mu_{n_k}) \geq I_{n_k}(\mu) \geq I(\mu). \tag{29}$$

By (A5), we have

$$g_{\mu_{n_i}}(x) \to g_{\mu_\infty}(x) \text{ uniformly on } \mathsf{X} \cap [-M, M]\,. \tag{30}$$

Hence $\langle \mu_{n_k} - \mu_{n_i},\ g_{\mu_{n_i}} \rangle \to 0$, as $i, k \to \infty$.

The function I is a continuous function on $\mathcal{M}(\sigma_P^2, M, \mathsf{X})$, by Proposition 2.2. Hence, from (29) and (30) we obtain

$$I(\mu_\infty) \geq I(\mu), \text{ for any } \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}). \tag{31}$$

So $\mu_\infty = \mu^*$ is optimal and both $I(\mu_n)$ and $I_n(\mu_n)$ converge to $I(\mu^*)$.

The proof of part (iv) is similar to [2, Thm. 2.2, 2.5], and can be found in other linear programming books as well . ☐

***Example*: The real AWGN channel**

Figures 5 and 6 show results from one experiment for the real AWGN example, where $\sigma_P^2 = 10$, $\sigma_N^2 = 10$ and $\mathsf{X} = [-10, -9, ..., 9, 10]$. The state space was taken finite to facilitate computation.

Some of these numerical results are surprising:

(i) When $M = \infty$, the optimal distribution is Gaussian. One might expect that μ^* on the twenty-one point state space X would approximate this continuous distribution. However, the results shown in Figure 6 show that the optimal distribution is supported on only five of these twenty-one points in X. On these five points however, the optimal distribution takes the form of a Gaussian distribution.

(ii) Figure 5 indicates that the optimal distribution μ^* on X very nearly achieves the same mutual information as the optimal Gaussian distribution on $\mathbb{R}$, even though μ^* is far simpler, with only 5 points of support.

(iii) It is found in this example that convergence of mutual information, as seen in Figure 5, is far faster than convergence of the distributions shown in Figure 6. This is again explained by the fact that a large family of distributions are nearly optimal for the AWGN channel with these parameters. ☐

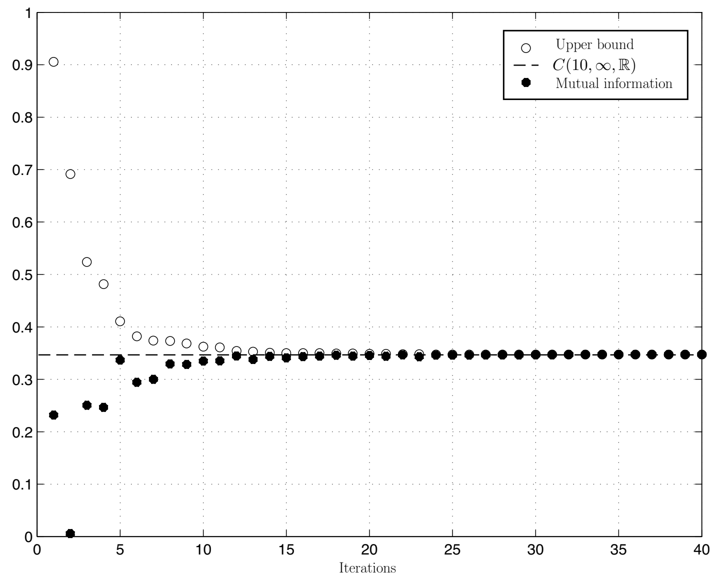

Figure 5: ***The real AWGN Channel*:** Convergence of the cutting-plane algorithm on $\mathcal{M}(10, 10, \mathsf{X})$.

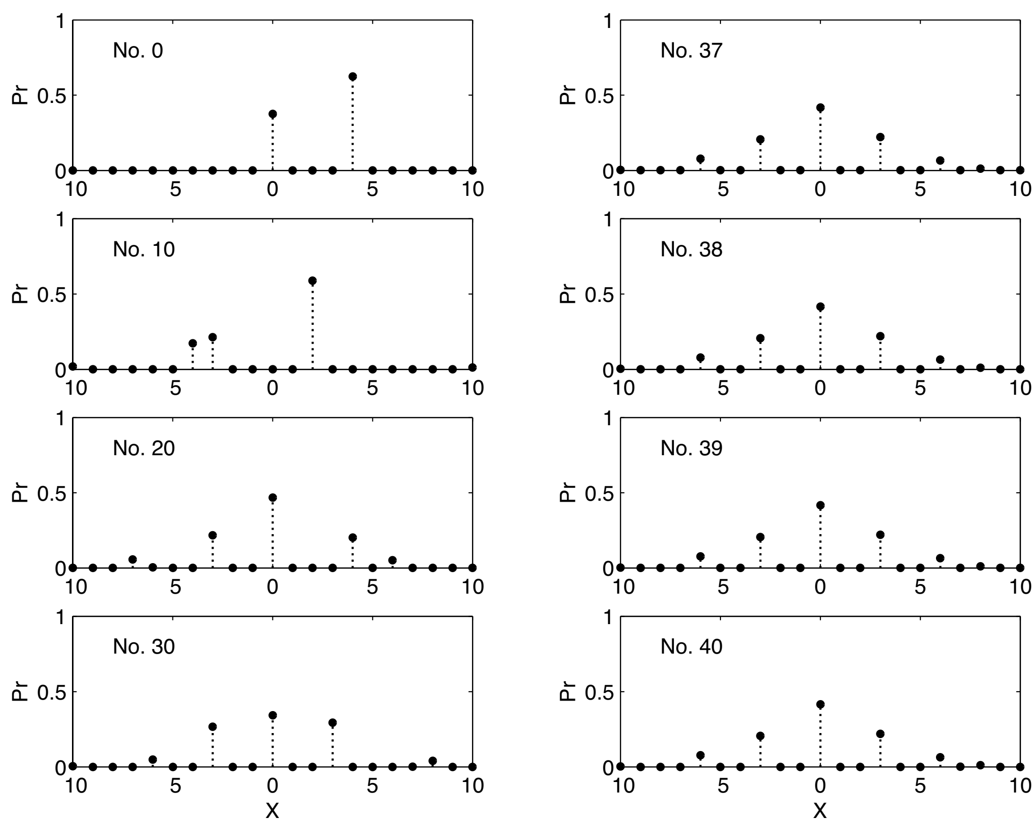

Figure 6: Convergence of input distributions for real AWGN channel.

***Example*: Complex channel models**

We conclude this subsection with several numerical examples for the complex AWGN, Rayleigh and Rician channels.

To facilitate computation we take X *finite* as in the previous example, so that the linear program (28) is finite dimensional for each n. We take

$$\mathsf{X} := \{0,1,2,3,4,5\}, \quad M = 5, \quad \sigma_A^2 = \sigma_N^2 = 1, \textit{ and } \quad \sigma_P^2 = 4\,. \tag{32}$$

In each case, the signal to noise ratio is 6 dB.

From the numerical results provided below we arrive at the following conclusions:

(i) In each experiment, for each of the three models, the mutual information $I(\mu_n)$, and the upper bound c_n converge rapidly to a common value.

(ii) In Figure 7 we find that the convergence is slowest for the AWGN channel, where the optimal input distribution shows greater complexity than seen in the Rician or Rayleigh channels.

(iii) In Figure 10 we see that the sequence of distributions $\{\mu_n\}$ obtained for the Rayleigh channel converges to a three-point distribution in just six iterations. Figure 12 shows that convergence is slightly slower for the Rician channel.

(iv) Generally, the convergence of the input distributions is slower than the convergence of mutual information. This suggests that the directional derivative of mutual information may be small in certain directions.

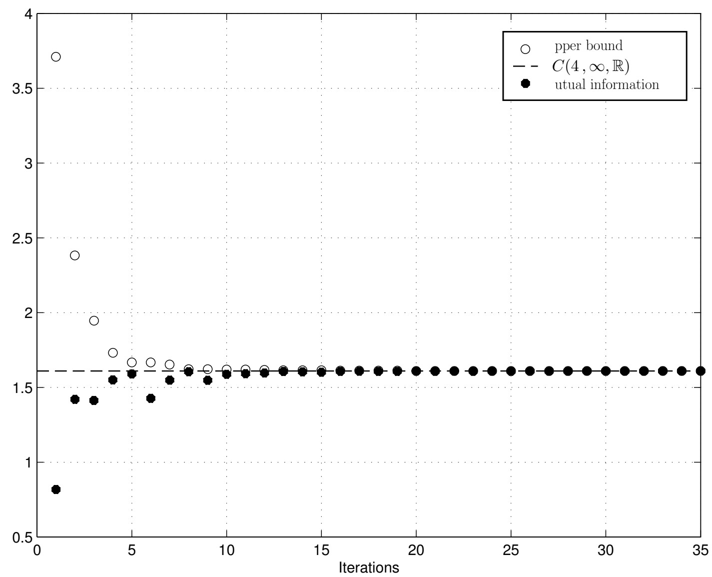

Figure 7: ***Complex AWGN Channel*:** Convergence of the cutting-plane algorithm on $\mathcal{M}(4, 5, \mathsf{X})$.

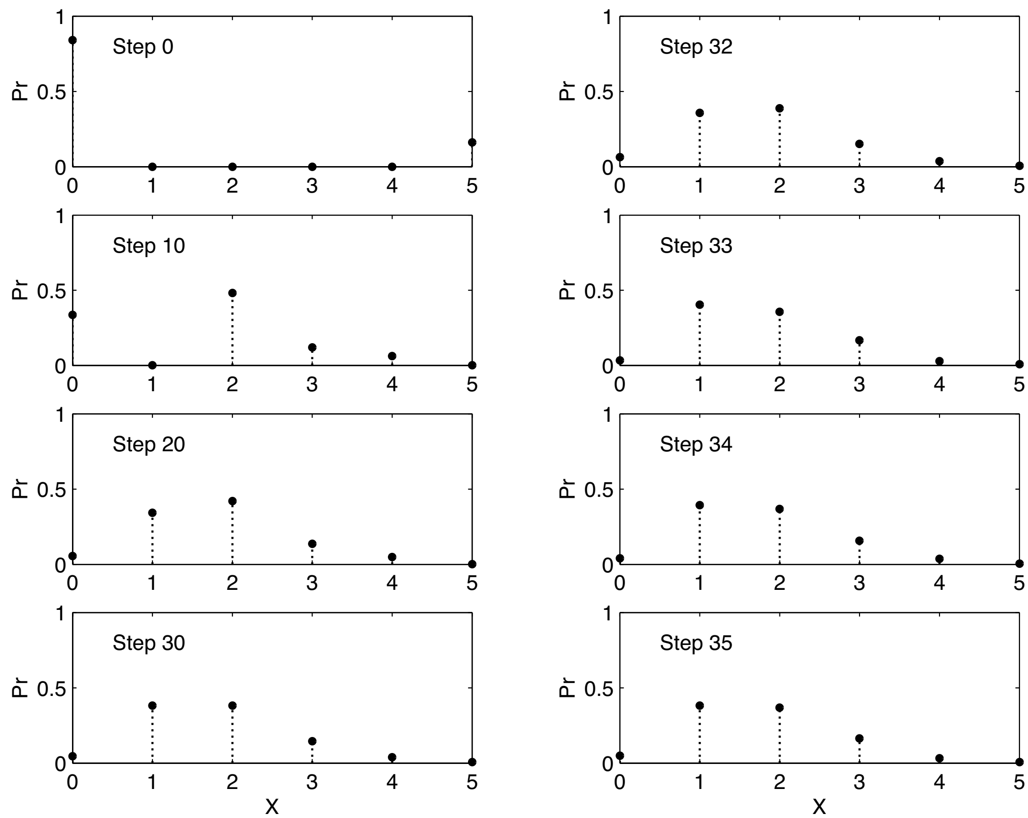

Figure 8: Convergence of input distributions for complex AWGN channel.

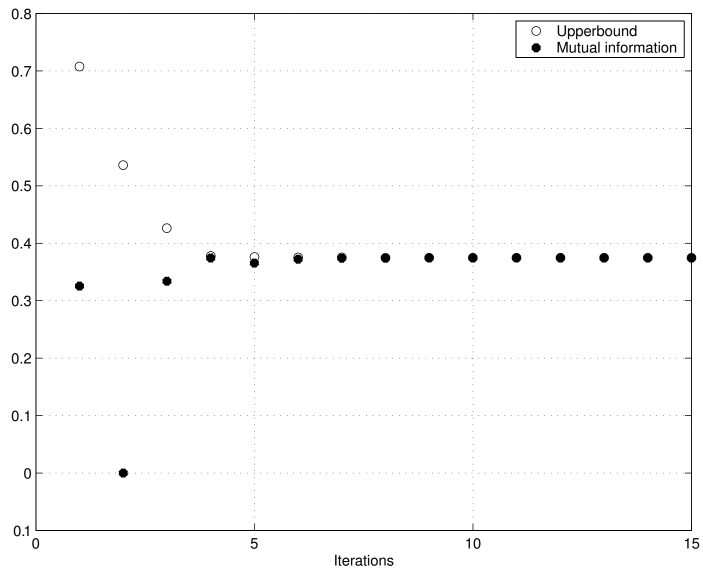

Figure 9: ***Rayleigh Channel*:** Convergence of the cutting-plane algorithm on $\mathcal{M}(4, 5, \mathsf{X})$.

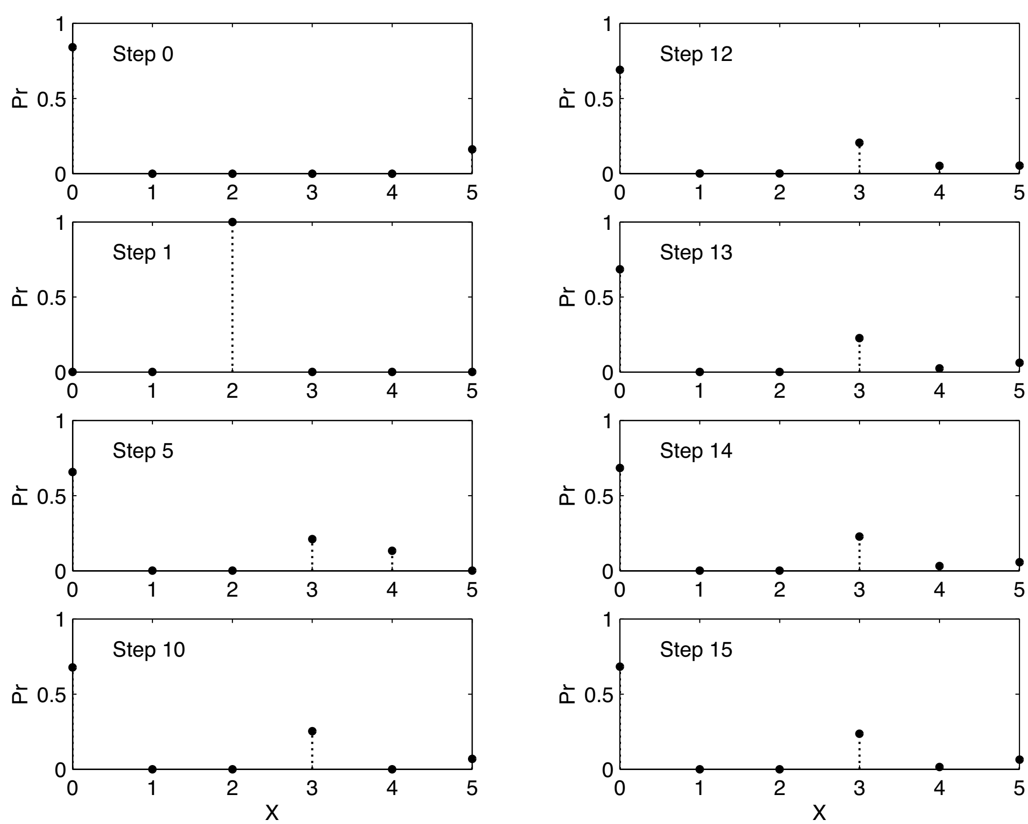

Figure 10: Convergence of input distributions for the Rayleigh channel.

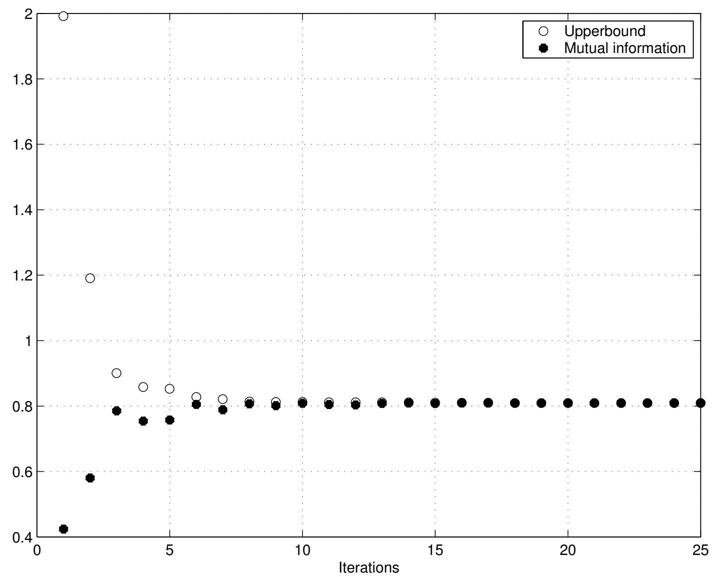

Figure 11: ***Rician Channel*****:** Convergence of the cutting-plane algorithm with constraint set $\mathcal{M}(4, 5, \mathsf{X})$.

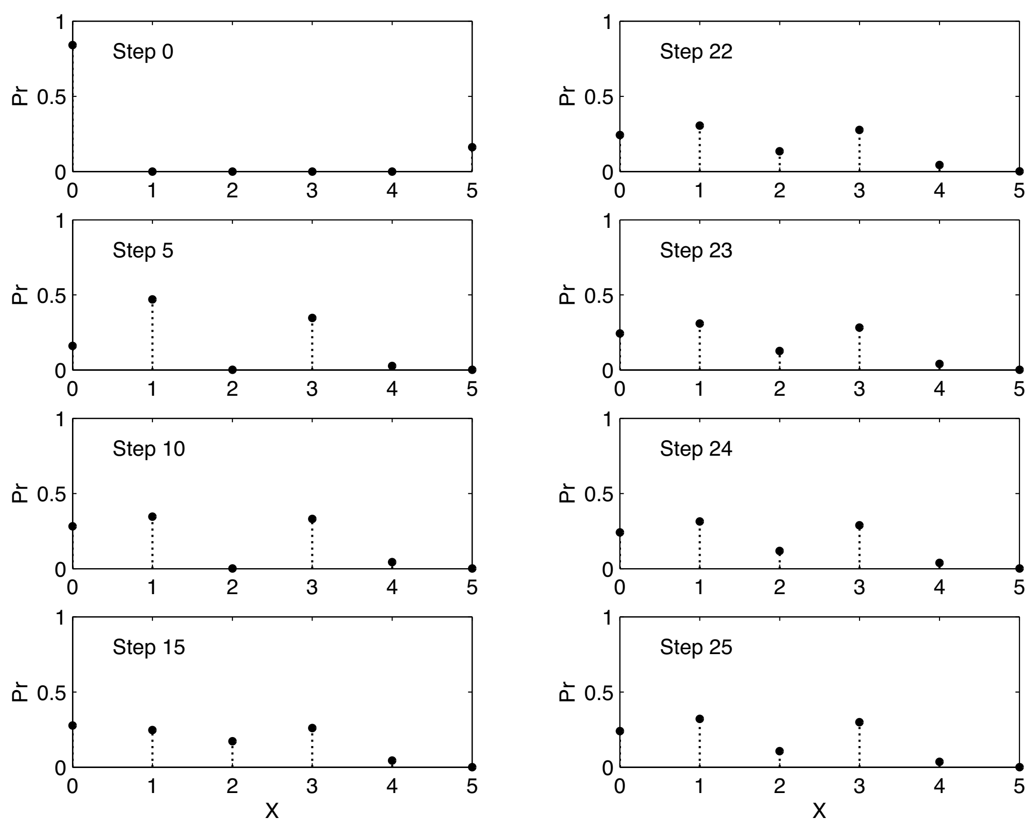

Figure 12: Convergence of input distributions for the Rician channel.

4.2 Computation of the optimal input alphabet

Although the cutting plane algorithm is convergent even in the infinite dimensional setting in which X is continuous, a finite dimensional algorithm is needed in any practical application. This is the reason that the input alphabet was taken to be fixed and finite in each of the numerical examples described in Section 4.1.

In this section we introduce an extension of the cutting plane method to *construct* the input alphabet X. Given an initial finite alphabet X_0, a sequence of finite alphabets $\{\mathsf{X}_n : n \geq 0\}$ is obtained by induction, each a subset of a closed interval $[-M, M]$. At the nth state of the algorithm, the optimal input distribution μ_n on X_n is obtained using the cutting plane algorithm introduced in Section 4.1. The details of this procedure are described as follows:

Steepest-ascent cutting plane algorithm

The algorithm is initialized with a finite alphabet X_0, together with a distribution $\mu_0 \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}_0)$. At the nth stage of the algorithm, we are given n distributions $\{\mu_0, \mu_2, \ldots, \mu_{n-1}\} \subset \mathcal{M}(\sigma_P^2, M, \mathsf{X})$, and an input alphabet X_{n-1}. The next distribution and input alphabet are then defined as follows:

(i) The nth piecewise linear approximation,

$$I_n(\mu) := \min_{0 \leq i \leq n-1} \langle g_{\mu_i}, \mu \rangle, \qquad \mu \in \mathcal{M}\,. \tag{33}$$

(ii) The next distribution,

$$\mu_n = \arg\max\{I_n(\mu) : \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})\}. \tag{34}$$

(iii) The new alphabet $\mathsf{X}_{n+1} = \mathsf{X}_n \cup \{x_{n+1}\}$, where

$$x_{n+1} = \arg\max\{g_n(x) - r_n x^2 : |x| \leq M, x \in \mathsf{X}\}, \tag{35}$$

$g_n(x) := g_{\mu_n}(x)$, and r_n is the associated Lagrange multiplier obtained in the solution of (34).

The algorithm is convergent for models with finite peak power constraint:

Theorem 4.2 *Suppose that (A1)–(A5) hold and $M < \infty$. Assume moreover that $\bar{F} := \sup_{\mu^1, \mu^2 \in \mathcal{M}} F(\mu^2; \mu^1) < \infty$. Then the steepest-ascent cutting plane algorithm has the following properties:*

(i) $I(\mu_n) \uparrow C(\sigma_P^2, M, \mathsf{X})$, *$n \to \infty$.*

(ii) $\mu_n \to \mu^*$ *weakly, as $n \to \infty$.*

(iii) $r_n \to r$, *where r is the Lagrange multiplier given in (22).*

Proof Let $L(\mu, r) := I(\mu) - r\langle \mu, \phi \rangle$, for $\mu \in \mathcal{M}$, $r \geq 0$, and for each $n \geq 0$ define two functions on the interval $[0, 1]$,

$$\begin{aligned} S_{n+1}(\theta) &:= L((1-\theta)\mu_n + \theta\delta_{x_{n+1}}, r_n) - L(\mu_n, r_n) \\ S^\circ_{n+1}(\theta) &:= L((1-\theta)\mu_n + \theta\mu_{n+1}, r_n) - L(\mu_n, r_n)\,, \qquad 0 \leq \theta \leq 1. \end{aligned}$$

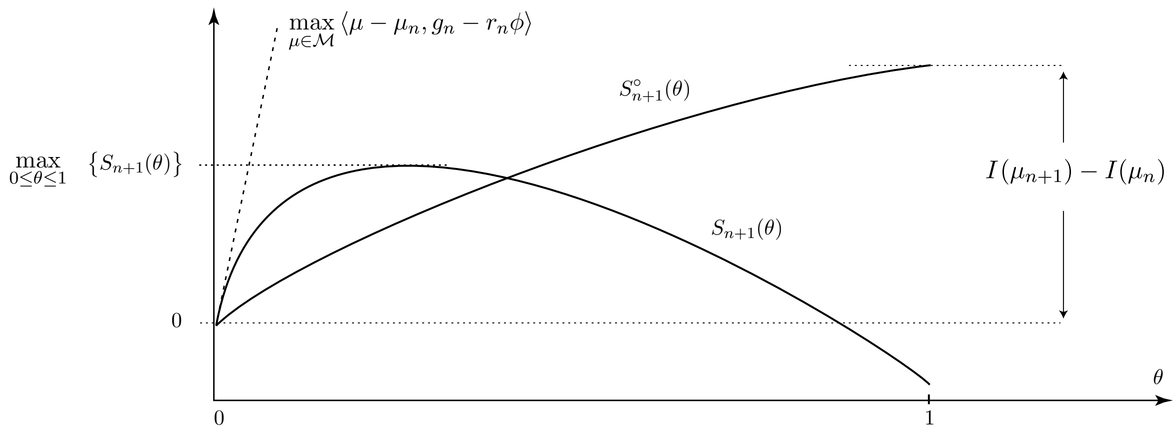

Figure 13: Convergence of the steepest-ascent cutting plane algorithm.

We set $\mathcal{E}_n = S'_n(0)$ for $n \geq 1$. From the definition of x_{n+1} and the derivative formula (18) we have,

$$\mathcal{E}_{n+1} = \langle \delta_{x_{n+1}} - \mu_n, g_n - r_n\phi \rangle = \sup_{\mu\in\mathcal{M}} \langle \mu - \mu_n, g_n - r_n\phi \rangle. \tag{36}$$

Applying Proposition 2.9 (ii), we see that to prove (i) it is enough to show that $\mathcal{E}_n \to 0$ as $n \to \infty$.

As shown in Figure 13, we have

$$\sup_{0\leq\theta\leq 1} S_{n+1}(\theta) \leq \varepsilon_{n+1} := I(\mu_{n+1}) - I(\mu_n)\,, \qquad n \geq 0. \tag{37}$$

Since $\{I(\mu_n) : n \geq 0\}$ is a bounded increasing sequence, it follows that ε_n is a summable sequence. In particular, $\varepsilon_n \to 0$ as $n \to \infty$. Consider then the Taylor series expression, for any given $\theta \in [0, 1]$,

$$S_{n+1}(\theta) = S_{n+1}(0) + \theta S'_{n+1}(0) + \tfrac{1}{2}\theta^2 S''_{n+1}(\tilde{\theta})$$

where $\tilde{\theta} \in (0, \theta)$. From the definitions and Proposition 2.6 we have

$$S_{n+1}(0) = 0\,, \quad S'_{n+1}(0) = \mathcal{E}_{n+1}\,, \quad \textit{and} \quad |S''_{n+1}(\tilde{\theta})| \leq \bar{F}.$$

Since we assume $\bar{F}$ is bounded, this combined with (37) gives the following bound,

$$\mathcal{E}_{n+1} \leq \theta^{-1}[\varepsilon_{n+1} + \bar{F}\theta^2]\,, \qquad 0 \leq \theta \leq 1.$$

The best bound is obtained on setting $\theta = \sqrt{\varepsilon_{n+1}/\bar{F}}$, which is less than one for sufficiently large n. We thus obtain, for $n \geq 0$ sufficiently large,

$$\mathcal{E}_{n+1} \leq \sqrt{\frac{\varepsilon_{n+1}}{\bar{F}}} \,.$$

This proves part (i).

To prove part (ii), suppose that $\{n_i\}$ is subsequence of $\{1, 2, 3, ...\}$ and that μ_∞ is a distribution on X such that

$$\mu_{n_i} \to \mu_\infty \text{ weakly, as } n \to \infty.$$

Then by part (i) and upper-semicontinuity of I, we have

$$C(\sigma_P^2, M, \mathsf{X}) = \limsup_{i\to\infty} I(\mu_{n_i}) \leq I(\mu_\infty).$$

By the uniqueness of μ^*, we obtain $\mu_\infty = \mu^*$. This proves part (ii). To prove (iii), note that from (36), we have

$$\langle \mu - \mu_n, g_n - r_n\phi \rangle \leq \mathcal{E}_{n+1} \,.$$

As $n \to \infty$, suppose $r_n \to r_\infty$, we have

$$\langle \mu - \mu^*, g_{\mu^*} - r_\infty\phi \rangle \leq 0,$$

that is

$$g_{\mu^*} \leq \Psi(r_\infty) + r_\infty\phi.$$

From the Kuhn-Tucker condition, we know $r_\infty = r$ is the Lagrange multiplier given in (22). □

Intuitively, the steepest-ascent cutting plane algorithm attempts to distort the peak of the sensitivity function g_n downwards so that it will fall below a quadratic function. In the special case where $\sigma_P^2 = \infty$, the algorithm attempts to impose an alignment condition between g_n and some constant function on $[-M, M]$, where the constant corresponds to channel capacity.

Observe that the size of the input alphabet grows linearly with n. Consequently, the complexity of each iteration grows since the cutting plane algorithm is more complex for larger input alphabets. However, if the algorithm is modified by discarding points with negligible probability, then the complexity of the algorithm may be bounded. This is illustrated in the following example.

***Example*: Steepest-ascent for the Rayleigh channel**

Recall that this is given by $V = AU + N$, with A and N mutually independent, circularly-symmetric complex Gaussian random variables. In our analysis and numerical results we consider the equivalent real channel whose transition density is shown in (13). The input and output are non-negative in the equivalent real channel, since each represents the scaled magnitude of their complex counterpart.

We first show that the conditions of Theorem 4.2 hold in this special case. Recall that properties (A1)-(A5) were established for the Rayleigh channel in Proposition 2.5. Hence it is enough to show that $\bar{F}$ is finite. This follows directly from (13):

$$\begin{aligned} \frac{p(y|\mu^2)}{p(y|\mu^1)} &= \frac{\int p(y|x)\mu^2(dx)}{\int p(y|x)\mu^1(dx)} \\ &\le \frac{\int \exp(\frac{-y}{1+M^2})\mu^1(dx)}{\int \frac{1}{1+M^2}\exp(-y)\mu^2(dx)}\,, \end{aligned}$$

which is bounded. Consequently, the algorithm is convergent by Theorem 4.2.

Figure 14 shows numerical results from one experiment using the steepest-accent cutting plane algorithm for this model with $\mathsf{X} = \mathbb{R}_+$, and the peak power constraint $|X| \le M = 10$. The algorithm was initialized with a binary distribution $\{0, 10\}$ with $0.2 = \mathsf{P}(X = 0) = 1 - \mathsf{P}(X = 10)$.

It appears from the figure that the algorithm converges to an optimal distribution satisfying the required alignment condition on the interval $[0, M]$ in just one iteration. The distribution μ^* obtained in Step 2 has only three points of support $\{0, x^*, 10\}$ where x^* is equal to the point at which g_{μ_0} achieves its maximum on $(0, 10)$.

In fact, at steps 3 and 4 there are respectively 2 and 3 points clustered around the point x^*. However, the mutual information is essentially unchanged in iterations 3 and 4. $\square$

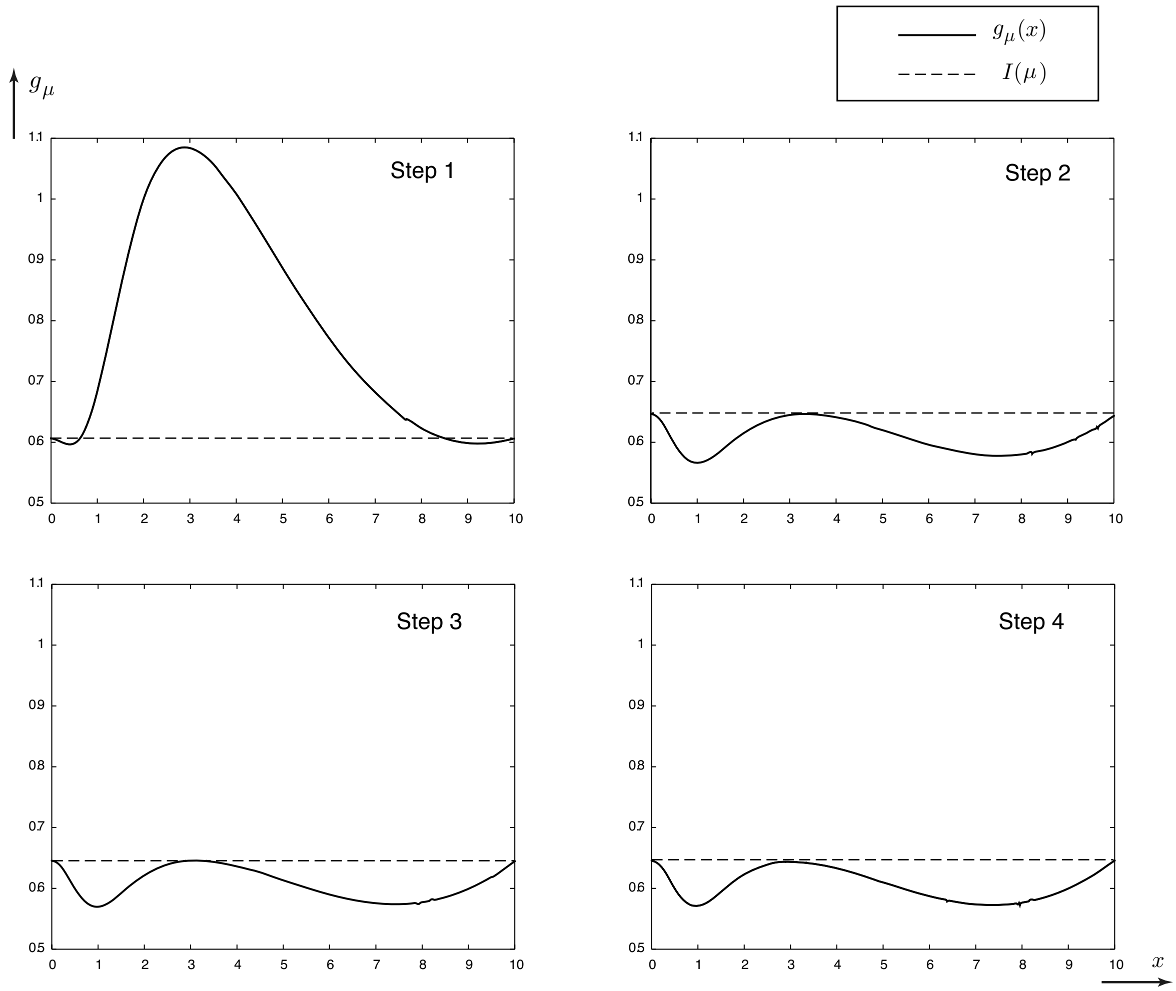

Figure 14: Convergence of the steepest-ascent cutting plane algorithm for the Rayleigh channel. It is seen that the alignment condition, and hence an optimal distribution is obtained after only two steps. The optimal distribution has three points of support, located at the points where g_μ reaches its upper-bound, as shown in Steps 2-4 above.

5 Conclusions

We have shown in this paper through theory and numerical results that in many instances it is possible to construct a simple, discrete distribution that performs nearly optimally in channel coding. Motivated by these findings, we have constructed a cutting plane algorithm and related algorithms to compute effective discrete distributions. We believe that these insights will provide new methodology for signal constellation design in a range of applications.

Several extensions are currently under investigation. In particular, we are considering multiple access models. In addition, we have recently discovered that the techniques introduced in this paper may be extended to optimization of the associated error exponent for a given capacity target. The duality theory developed in the recent papers [13, 11] strongly suggests that extensions of this algorithm will also be effective in lossy data compression.

Finally, with a deeper understanding of the sensitivity of mutual information with respect to various parameters we expect to achieve a deeper understanding of the impact on capacity by channel uncertainty, and channel variation. Such insights may also lead to refinements of the algorithm considered here. Bounding techniques such as those employed in [21] will likely prove useful in this analysis.

Acknowledgment Prof. M. Medard at MIT provided numerous suggestions for extending this research, as well as many useful references. The authors gratefully acknowledge this inspiration and support.

References

[1] I.C. Abou-Faycal, M.D. Trott, and S. Shamai. The capacity of discrete-time memoryless Rayleigh-fading channels. *IEEE Trans. Inform. Theory*, 47(4):1290–1301, 2001.

[2] E.J. Anderson and P. Nash. *Linear programming in infinite-dimensional spaces.* Wiley, New York, 1987.

[3] D.P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, Massachusetts, 1999.

[4] E. Biglieri, J. Proakis, and S. Shamai. Fading channels: information-theoretic and communications aspects. *IEEE Trans. Inform. Theory*, 44(6):2619–2692, 1998.

[5] P. Billingsley. *Convergence of probability measures.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[6] R.E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, 18(4):460–473, 1972.

[7] R.E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Inform. Theory*, 20(4):405–417, 1974.

[8] R.E Blahut. *Principles and Practice of Information Theory.* McGraw-Hill, New York, 1995.

[9] S. Boyd and L. Vandenberghe. Convex optimization. Monograph available on-line at `http://www.stanford.edu/ boyd/cvxbook.html`, 2003.

[10] R. Chen, B. Hajek, and R. Koetter. On fixed input distributions for noncoherent communication over high SNR Rayleigh fading channels. *Submitted to IEEE Trans. Inform. Theory*, 2002.

[11] M. Chiang and S. Boyd. Geometric programming duals of channel capacity and rate distortion. Submitted to *IEEE Trans. Inform. Theory*, 2002.

[12] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley, New York, 1991.

[13] T. M. Cover and M. Chiang. Duality between channel capacity and rate distortion with two-sided state information. *IEEE Trans. Inform. Theory*, 48(6):1629–1638, 2002. Special issue on Shannon theory: perspective, trends, and applications.

[14] R. Gallager. Power limited channels: coding, multiaccess, and spread spectrum. In *Proc. 1988 Conf. Inform. Sci. Syst.*, page 372, 1988.

[15] R.G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[16] J.-L. Goffin and J.-P. Vial. Convex nondifferentiable optimization: a survey focused on the analytic center cutting plane method. *Optim. Methods Softw.*, 17(5):805–867, 2002.

[17] J. Huang and S. Meyn. Characterization and computation of optimal distributions for channel coding. In *Proceedings of the 37th Annual Conference on Information Sciences and Systems (CISS)*, March 2003.

[18] J.E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

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

[20] T.L. Marzetta and B.M. Hochwald. Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading. *IEEE Trans. Inform. Theory*, 45(1):139–157, 1999.

[21] M. Medard. The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel. *IEEE Trans. Inform. Theory*, 46(3):933–946, 2000.

[22] R. Palanki. On the capacity-achieving distributions of some fading channels. In *Proceedings of 40th Allerton Conference on Communication, Control, and Computing*, 2002.

[23] C. Pandit and S. Meyn. Robust hypothesis testing with application to sensor networks. In preparation. Preliminary report presented at 2003 IEEE International Symposium on Information Theory, Yokohama, JAPAN June 29 - July 4, 2003.

[24] C. Pandit and S. Meyn. A novel algorithm for measurement-based admission control using worst-case source characterizations. Submitted to *IEEE/ACM Transactions on Networking*. Preliminary report published in the *Proceedings of the 37th Annual Conference on Information Sciences and Systems* (CISS), 2003, August 2003.

[25] J.G Proakis. *Digital Communications*. McGraw-Hill, New York, 1995.

[26] J.S. Richters. Communication over fading dispersive channels. Technical Report 464, MIT Res. Lab Electron., Nov. 30, 1967.

[27] S. Shamai and I. Bar-David. The capacity of average and peak-power-limited quadrature Gaussian channels. *IEEE Trans. Inform. Theory*, 41(4):1060–1071, 1995.

[28] S. Shamai and S. Verdu. Worst-case power-constrained noise for binary-input channels. *IEEE Trans. Inform. Theory*, 38(5):1494–1511, 1992.

[29] C.E. Shannon and W. Weaver. *A Mathematical Theory of Communication*. Univ. of Illinois Press, Urbana, 1949.

[30] J.G. Smith. The information capacity of amplitude and variance-constrained scalar Gaussian channels. *Inform. Contr.*, 18:203–219, 1971.

[31] S. Verdu. On channel capacity per unit cost. *IEEE Trans. Inform. Theory*, 36(5):1019–1030, 1990.