

# ENTROPY, INFERENCE, AND CHANNEL CODING

J. HUANG\*, C. PANDIT†, S.P. MEYN‡, M. MÉDARD§, AND V. VEERAVALLI¶

**Abstract.** This article surveys application of convex optimization theory to topics in Information Theory. Topics include optimal robust algorithms for hypothesis testing; a fresh look at the relationships between channel coding and robust hypothesis testing; and the structure of optimal input distributions in channel coding.

A key finding is that the optimal distribution achieving channel capacity is typically discrete, and that the distribution achieving an optimal error exponent for rates below capacity is *always* discrete. We find that the resulting codes significantly out-perform traditional signal constellation schemes such as QAM and PSK.

**AMS(MOS) subject classifications.** Primary: 94A24, 94A13, 94A17, Secondary: 94A34, 94A40, 60F10

**Key words.** Information theory; channel coding; error exponents; fading channels.

**1. Introduction.** This article surveys application of convex optimization theory to topics in information theory. Our main focus is on channel coding, and the relationships between channel coding and robust hypothesis testing.

The optimization problems considered in this paper concern minimization or maximization of a convex function over the space of probability measures. The focus is on the following three central areas of information theory: hypothesis testing, channel capacity, and the exponential bounds on the probability of error in channel coding. One foundation of this work lies in the theory of convex optimization [5, 9]. In particular, the structural properties obtained are based on convex duality theory and the Kuhn-Tucker alignment conditions. A second foundation is entropy. Recall that for two distributions  $\mu, \pi$  the relative entropy, or Kullback-Leibler divergence is defined as,

$$D(\mu \parallel \pi) = \begin{cases} \int \log\left(\frac{\mu(dx)}{\pi(dx)}\right) \mu(dx) & \text{if } \mu \prec \pi, \\ \infty & \text{otherwise} \end{cases}$$

Relative entropy plays a fundamental role in hypothesis testing and communications, and it arises as the natural answer to several important ques-

---

\*Marvell Technology, Santa Clara, CA. [jianyh@marvell.com](mailto:jianyh@marvell.com)

†Morgan Stanley and Co., 1585 Broadway, New York, NY, 10019  
[charuhas.pandit@morganstanley.com](mailto:charuhas.pandit@morganstanley.com)

‡Department of Electrical and Computer Engineering and the Coordinated Sciences Laboratory, University of Illinois at Urbana-Champaign [meyn@uiuc.edu](mailto:meyn@uiuc.edu). Supported in part by NSF grant ITR 00-85929.

§Laboratory for Information and Decision Systems, Massachusetts Institute of Technology [medard@mit.edu](mailto:medard@mit.edu). Supported in part by NSF grant Career 6891730.

¶Department of Electrical and Computer Engineering and the Coordinated Sciences Laboratory, University of Illinois at Urbana-Champaign [vvv@uiuc.edu](mailto:vvv@uiuc.edu). Supported in part by NSF grant ITR 00-85929.

tions in applications in data compression, model-selection in statistics, and signal processing [29, 15, 12, 4, 16, 13, 14, 20, 36, 8, 17, 31].

**1.1. Channel models.** We consider a stationary, memoryless channel with input alphabet  $\mathsf{X}$ , output alphabet  $\mathsf{Y}$ , and transition density defined by

$$P(Y \in dy \mid X = x) = p(y|x) dy, \quad x \in \mathsf{X}, y \in \mathsf{Y}. \quad (1.1)$$

It is assumed that  $\mathsf{Y}$  is equal to either  $\mathbb{R}$  or  $\mathbb{C}$ , and we assume that  $\mathsf{X}$  is a closed subset of  $\mathbb{R}$ . For a given input distribution  $\mu$  on  $\mathsf{X}$ , the density for the marginal distribution of the output is denoted,

$$p_\mu(dy) = \int \mu(dx)p(y|x), \quad y \in \mathsf{Y}. \quad (1.2)$$

Many complex channel models in which  $\mathsf{X}$  is equal to  $\mathbb{C}$  are considered by viewing  $\mu$  as the amplitude of  $X$ . Details on this transformation are given prior to Theorem 1.1 below.

Throughout the paper we restrict to *noncoherent* channels in which neither the transmitter nor the receiver knows the channel state.

A *signal constellation* is a finite set of points in  $\mathsf{X}$  that is used to define possible codewords. Two well known examples when  $\mathsf{X} = \mathbb{C}$  are *quadrature-amplitude modulation* (QAM), and *phase-shift keyed* (PSK) coding. In these coding schemes the codewords are chosen using a random code constructed with a uniform distribution across the given signal constellation. These methods are largely motivated by properties of the additive Gaussian noise (AWGN) channel, where it is known that a random code book obtained using a Gaussian distribution achieves capacity.

The problem of constellation design has recently received renewed attention in information theory and communication theory. While many techniques in information theory such as coding have readily found their way into communication applications, the signal constellations that information theory envisages and those generally considered by practitioners differ significantly. In particular, while the optimum constellation for an AWGN channel is a continuous constellation that allows for a Gaussian distribution on the input, commonly used constellations over AWGN channels, such as quadrature amplitude modulation (QAM), are not only discrete, but also generally regularly spaced. This gap between theory and practice can be explained in part by the difficulty of deploying, in practical systems, continuous constellations.

However, there is also a body of work which strongly suggests the continuous paradigm favored by theoreticians is inappropriate for realistic channel models in the majority of today's applications, such as wireless communication systems. Under any of the following conditions the optimal capacity achieving distribution has a finite number of mass points, or in the case of a complex channel, the amplitude has finite support:

- (i) The AWGN channel under a peak power constraint [38, 37, 32, 10].
- (ii) Channels with fading, such as Rayleigh [1] and Rician fading [22, 21]. Substantial generalizations are given in [25].
- (iii) Lack of channel coherence [27]. For the noncoherent Rayleigh fading channel, a Gaussian input is shown to generate bounded mutual information as SNR goes to infinity [11, 30].
- (iv) Under general conditions a binary distribution is optimal, or approximately optimal for sufficiently low SNR ([19], and [39, Theorem 3].)

This article provides theory to explain why optimal distributions are typically discrete based on the Kuhn-Tucker alignment conditions.

**1.2. Capacity and error exponents.** Operator-theoretic notation is convenient in the convex-analytic setting of this article. We let  $\mathcal{M}$  denote the set of probability distributions on the Borel sigma field on some state space  $\mathsf{X}$ , which is always taken to be a closed subset of Euclidean space. For any distribution  $\mu \in \mathcal{M}$ , and any measurable function  $f: \mathsf{X} \rightarrow \mathbb{R}$ , we denote

$$\langle \mu, f \rangle := \int f(x) \mu(dx).$$

Mutual information is defined as,

$$I(\mu) = \int \left( \int \log \left( \frac{p(y|x)}{p_\mu(y)} \right) p(y|x) dy \right) \mu(dx), \quad \mu \in \mathcal{M}, \quad (1.3)$$

and channel capacity is determined by maximizing mutual information subject to two linear constraints:

- (i) The *average power constraint* that

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

where  $\phi(x) := x^2$  for  $x \in \mathbb{R}$ .

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

Hence the input distribution is constrained by  $\mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$ , where

$$\mathcal{M}(\sigma_P^2, M, \mathsf{X}) := \left\{ \mu \in \mathcal{M} : \langle \mu, \phi \rangle \leq \sigma_P^2, \mu\{[-M, M]\} = 1 \right\}, \quad (1.4)$$

and the capacity  $C(\sigma_P^2, M, \mathsf{X})$  is expressed as the value of a convex program,

$$\begin{aligned} & \mathbf{sup} && I(\mu) \\ & \mathbf{subject\ to} && \mu \in \mathcal{M}(\sigma_P^2, M, \mathsf{X}). \end{aligned} \quad (1.5)$$

The *channel reliability function* is denoted,

$$E(R) = \lim_{N \rightarrow \infty} \left[ -\frac{1}{N} \log p_e(N, R) \right], \quad R > 0, \quad (1.6)$$

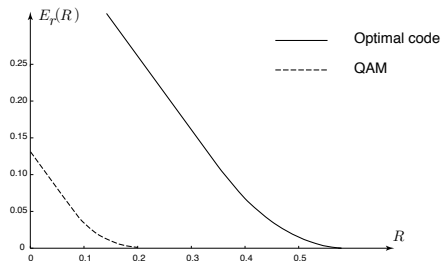


Fig. 1: The error exponent  $E_r(R)$  for the two codes shown in Figure 11. The 3-point constellation performs better than 16-point QAM for all rates  $R \leq C$ .

where  $p_e(N, R)$  is the minimal probability of error, where the minimum is over all *block codes* of length  $N$  and rate  $R$ . The *random coding exponent*  $E_r(R)$  may be expressed for a given  $R < C$  via,

$$E_r(R) = \sup_{0 \leq \rho \leq 1} \left( \sup_{\mu \in \mathcal{M}(\sigma_P^2, M, \mathbf{X})} \left[ -\rho R - \log(G^\rho(\mu)) \right] \right), \quad (1.7)$$

where for each  $\rho \geq 0$  we define,

$$G^\rho(\mu) := \int \left[ \int \mu(dx) p(y|x)^{1/(1+\rho)} \right]^{1+\rho} dy. \quad (1.8)$$

The following *random-coding bound* holds under the assumptions imposed in this paper,

$$p_e(N, R) \leq \exp(-NE_r(R)), \quad N \geq 1, R \geq 0.$$

Moreover, the equality  $E(R) = E_r(R)$  holds for rates greater than the *critical rate*  $R_{\text{crit}}$  [7, 18].

If one can design a distribution with a large error exponent, then the associated random code can be constructed with a correspondingly small block-length. This has tremendous benefit in implementation. Figure 1 shows an example taken from [26] that illustrates how a better designed code can greatly out-perform QAM for rates  $R$  below capacity.

Optimization of the random coding exponent is addressed as follows. Rather than parameterize the optimization problem by the given rate  $R > 0$ , we consider for each Lagrange multiplier  $\rho$  the convex program,

$$\begin{aligned} & \mathbf{inf} && G^\rho(\mu) \\ & \mathbf{subject\ to} && \mu \in \mathcal{M}(\sigma_P^2, M, \mathbf{X}). \end{aligned} \quad (1.9)$$

The value is denoted  $G^{\rho*}$ , and we have from (1.7)

$$E_r(R) = \sup_{0 \leq \rho \leq 1} [-\rho R - \log(G^{\rho*})].$$

Our objective is to explore the structure of optimal input distributions achieving either channel capacity or the random coding exponent  $E_r(R)$ . Instead of studying individual channel models, which have been the topics of numerous papers, we take a systematic approach to study these problems under very general channel conditions. We believe this viewpoint will clarify the our applications of optimization theory to information theory.

**1.3. Assumptions and examples.** The basis of our analysis is the structure of two *sensitivity functions* that may be interpreted as gradients with respect to  $\mu$  of the respective objective functions  $I(\mu)$  and  $G^\rho(\mu)$ . The *channel sensitivity function* is defined by

$$g_\mu(x) := \int \log[p(y|x)/p_\mu(y)]p(y|x) dy, \quad x \in \mathbb{R}, \quad (1.10)$$

where  $p_\mu$  was defined in (1.2). For each  $x$ ,  $g_\mu(x)$  is the relative entropy between two probability distributions  $p(y|x)$  and  $p_\mu(y)$ . The *error exponent sensitivity function* is given by,

$$g_\mu^\rho(x) := \int \left[ \int \mu(dz) p(y|z)^{1/(1+\rho)} \right]^\rho p(y|x)^{1/(1+\rho)} dy, \quad x \in \mathbb{X}. \quad (1.11)$$

The following limit follows from elementary calculus,

$$-g_\mu(x) = \lim_{\rho \rightarrow 0} \frac{\log g_\mu^\rho(x)}{\rho}, \quad x \in \mathbb{R}.$$

The existence of a solution to (1.5) or (1.9) 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  $\mathbb{X}$  is a closed subset of  $\mathbb{R}$ ,  $\mathbb{Y} = \mathbb{C}$  or  $\mathbb{R}$ , and  $\min(\sigma_P^2, M) < \infty$ .

(A2) For each  $n \geq 1$ ,

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

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

Conditions (A1)-(A3) are also the standing conditions in [25, 26].

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  $U \in \mathbb{U} = \text{a closed subset of } \mathbb{C}$ , and  $V \in \mathbb{V} = \mathbb{C}$ . The input and output are related by the transition density via,

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

The optimization problem (1.5) is unchanged: The average power constraint is given by  $\mathbb{E}[|U|^2] \leq \sigma_p^2$ , and the peak-power constraint indicates that  $|U| \leq M$ , where  $|z|$  denotes the modulus of a complex number  $z \in \mathbb{C}$ . It is always assumed that the complex model is symmetric: *symmetric*:

$$p_{\bullet}(v|u) = p_{\bullet}(e^{j\alpha}v|e^{j\alpha}u), \quad u, v \in \mathbb{C}, \alpha \in \mathbb{R}. \quad (1.12)$$

Under (1.12) we define,

- (i)  $X = |U|$ ,  $\mathsf{X} = \mathsf{U} \cap \mathbb{R}_+$ , and  $\mathcal{M}$  again 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  $\mu$ . That is, we have the polar-coordinates representation,

$$\mu_{\bullet}(dx \times d\alpha) = \frac{1}{2\pi x} \mu(dx) d\alpha, \quad x > 0, 0 \leq \alpha \leq 2\pi, \quad (1.13)$$

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

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

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

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

where  $\mu_{\bullet}$  and  $\mu$  correspond as in (ii).

The symmetry condition (1.12) is a natural assumption in many applications since phase information is lost at high bandwidths. It is shown in [25] that in all of the standard complex channel models, Assumptions (A1)-(A3) hold for the corresponding real channel with input  $X = |U|$ . Moreover, both the capacity and random coding exponent  $E_r(R)$  for the complex and real models coincide.

We recall the most common channel models here:

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

$$V = (A + a)U + N \quad (1.14)$$

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  $N \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_N^2)$ . The Rician channel reduces to the complex AWGN channel when  $\sigma_A^2 = 0$ .

Throughout this 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.  $\square$

On setting  $a = 0$  in (1.14) we obtain another important special case:

**The Rayleigh channel** The model (1.14) 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 [2]: Setting  $X = |U|_{\sigma_A}/\sigma_N$  and  $Y = |V|^2/\sigma_N^2$ , we obtain a real channel model with transition density

$$p(y|x) = \frac{1}{1+x^2} \exp\left(-\frac{1}{1+x^2} y\right), \quad x, y \in \mathbb{R}_+. \quad (1.15)$$

□

**The phase-noise channel** This noncoherent AWGN channel emerges in communication systems where it is not possible to provide a carrier phase reference at the receiver [28]. The channel model is represented by

$$V = Ue^{j\theta} + N,$$

where  $U$  and  $V$  are the complex-valued channel input and output,  $N$  is an independent complex Gaussian random variable with variance  $2\sigma_N^2$ , and  $\theta$  is an independent random phase distributed uniformly on  $[-\pi, \pi]$ . It is easy to see the input phase does not convey any information, and the mutual information is decided by the conditional probability density of the channel output amplitude  $Y$  given the channel input magnitude  $X$ ,

$$p(y|x) = \frac{y}{\sigma_N^2} \exp\left(-\frac{y^2 + x^2}{2\sigma_N^2}\right) I_0\left(\frac{xy}{\sigma_N^2}\right), \quad (1.16)$$

where  $I_0$  is the zeroth-order modified Bessel function of the first kind. □

The sensitivity function  $g_\mu$  is easily computed numerically for the Rayleigh or phase-noise channels based on (1.15) or (1.16). For the general Rician model, computation of  $g_\mu$  appears to be less straightforward since this requires computation of  $g_{\mu_\bullet}$ , which involves integration over the complex plane.

The capacity-achieving input distribution is discrete under conditions imposed here when  $M$  is finite [25]. We find that the distribution optimizing the error exponent  $E_r$  for a given positive rate  $R < C$  *always* has finite support, with or without a peak power constraint. Consequently, in the symmetric complex channel, the distribution is symmetric, and its magnitude has finite support.

The following result provides a summary of results obtained for the real channel or the symmetric complex channel. The proof of Theorem 1.1 (ii) follows directly from Propositions 3.2 and 3.3.

**THEOREM 1.1.** *The following hold for the real channel model under Assumptions (A1)–(A3):*

- (i) *If  $M < \infty$  then there exists an optimizer  $\mu^*$  of the convex program (1.5) defining capacity. The distribution  $\mu^*$  has finite support.*

- (ii) Consider the convex program (1.9) under the relaxed condition that  $\min(\sigma_P^2, M) < \infty$ . For each  $\rho$  there exists an optimizer  $\mu^\rho$ , and any optimizer has finite support. Moreover, for each  $R \in (0, C)$  there exists  $\rho^*$  achieving the maximum in (1.7) so that,

$$E_r(R) = -\rho^* R - \log(G^{\rho^*}) = -\rho^* R - \log(G^{\rho^*}(\mu^{\rho^*})).$$

□

The remainder of the paper is organized as follows:

Section 2 reviews well known results on hypothesis testing, along with a promising new approach to robust hypothesis testing. The formulae for capacity and the random coding exponent are explained, based on results from robust and ordinary hypothesis testing.

Section 3 contains results from [25, 26] showing why the capacity-achieving input distribution and the distribution optimizing the error exponent are discrete. The numerical results contained in Section 4 show that the resulting codes can *significantly* out-perform traditional signal constellation schemes such as QAM and PSK. Section 5 concludes the paper.

**2. Hypothesis testing and reliable communication.** The focus of this section is to characterize optimal input distributions in hypothesis testing, channel capacity, and error exponents. A goal is to clarify the relationship between the solutions to these three optimization problems.

In Section 2.1 we survey some results from [23, 6, 41] on asymptotic hypothesis testing based on Sanov's Theorem. These results will be used to set the stage for the convex analytic methods and geometric intuition to be applied in the remainder of the paper.

We first briefly recall Sanov's Theorem: If  $\mathbf{X}$  is a real-valued sequence, the empirical distributions are defined as the sequence of discrete probability distributions on  $\mathcal{B}$ ,

$$\Gamma_N(A) = \frac{1}{N} \sum_{k=0}^{N-1} \mathbb{I}\{X_k \in A\}, \quad A \in \mathcal{B}. \quad (2.1)$$

Suppose that  $\mathbf{X}$  is i.i.d. with marginal distribution  $\pi$ . Sanov's Theorem states that for any closed convex set  $\mathcal{A} \subseteq \mathcal{M}$ ,

$$\lim_{N \rightarrow \infty} -N^{-1} \log \mathbf{P}\{\Gamma_N \in \mathcal{A}\} = \inf\{D(\mu \parallel \pi) : \mu \in \mathcal{A}\}.$$

The relative entropy is jointly convex on  $\mathcal{M} \times \mathcal{M}$ , and hence computation of the minimum of  $D(\mu \parallel \pi)$  amounts to solving a convex program.

**2.1. Neyman-Pearson hypothesis testing.** Consider the binary hypothesis testing problem based on a finite number of observations from a sequence  $\mathbf{X} = \{X_t : t = 1, \dots\}$ , taking values in the set  $\mathbf{X} = \mathbb{R}^d$ . It is assumed that, conditioned on the hypotheses  $H_0$  or  $H_1$ , these observations

are independent and identically distributed (i.i.d.). The marginal probability distribution on  $\mathbf{X}$  is denoted  $\pi^j$  under hypothesis  $H_j$  for  $j = 0, 1$ . The goal is to classify a given set of observations into one of the two hypotheses.

For a given  $N \geq 1$ , suppose that a decision test  $\phi_N$  is constructed based on the finite set of measurements  $\{X_1, \dots, X_N\}$ . This may be expressed as the characteristic function of a subset  $A_1^N \subset \mathbf{X}^N$ . The test declares that hypothesis  $H_1$  is true if  $\phi_N = 1$ , or equivalently,  $(X_1, X_2, \dots, X_N) \in A_1^N$ . The performance of a *sequence* of tests  $\phi := \{\phi_N : N \geq 1\}$  is reflected in the error exponents for the type-II error probability and type-I error probability, defined respectively by,

$$I_\phi := -\liminf_{N \rightarrow \infty} \frac{1}{N} \log(\mathbb{P}_{\pi^1}(\phi_N(X_1, \dots, X_N) = 0)),$$

$$J_\phi := -\liminf_{N \rightarrow \infty} \frac{1}{N} \log(\mathbb{P}_{\pi^0}(\phi_N(X_1, \dots, X_N) = 1)),$$

The asymptotic N-P criterion of Hoeffding [23] is described as follows: For a given constant  $\eta \geq 0$ , an optimal test is the solution to the following optimization problem,

$$\sup_{\phi} I_\phi \quad \text{subject to} \quad J_\phi \geq \eta \quad (2.2)$$

where the supremum is over all test sequences  $\phi$ .

The optimal value of the exponent  $I_\phi$  in the asymptotic N-P problem is described in terms of relative entropy. It is shown in [41] that one may restrict to tests of the following form without loss of generality: for a closed set  $\mathcal{A} \subseteq \mathcal{M}$ ,

$$\phi_N = \mathbb{I}\{\Gamma_N \in \mathcal{A}\}, \quad (2.3)$$

where  $\{\Gamma_N\}$  denotes the sequence of empirical distributions (2.1). Sanov's Theorem tells us that for any test of this form,

$$I_\phi = \inf\{D(\gamma \parallel \pi^1) : \gamma \in \mathcal{A}^c\}, \quad J_\phi = \inf\{D(\gamma \parallel \pi^0) : \gamma \in \mathcal{A}\}.$$

For an arbitrary measure  $\pi \in \mathcal{M}$  and for  $\beta \in \mathbb{R}_+$ , consider the *divergence set*,

$$\mathcal{Q}_\beta^+(\pi) := \{\gamma \in \mathcal{M} : D(\gamma \parallel \pi) \leq \beta\}. \quad (2.4)$$

The divergence set  $\mathcal{Q}_\beta^+(\pi)$  is a closed convex subset of  $\mathcal{M}$  since  $D(\cdot \parallel \cdot)$  is jointly convex and lower semi-continuous on  $\mathcal{M} \times \mathcal{M}$ . Consequently, from Sanov's Theorem the smallest set  $\mathcal{A}$  that gives  $I_\phi \geq \eta$  is the divergence set  $\mathcal{A}^* = \mathcal{Q}_\eta^+(\pi^0)$ , and hence the solution to (2.2) is the value of the convex program,

$$\beta^* = \sup\{\beta \geq 0 : \mathcal{Q}_\eta(\pi^0) \cap \mathcal{Q}_\beta(\pi^1) = \emptyset\} = \inf_{\gamma \in \mathcal{Q}_\eta(\pi^0)} D(\gamma \parallel \pi^1). \quad (2.5)$$

Theorem 2.1 may be interpreted geometrically as follows. We have  $\gamma^* \in \mathcal{Q}_\eta^+(\pi^0) \cap \mathcal{Q}_{\beta^*}^+(\pi^1)$ , and the convex sets  $\mathcal{Q}_\eta^+(\pi^0)$  and  $\mathcal{Q}_{\beta^*}^+(\pi^1)$  are separated by the following set, which corresponds to the test sequence in (2.7):

$$\mathcal{H} = \{\gamma \in \mathcal{M} : \langle \gamma, \log \ell \rangle = \langle \gamma^*, \log \ell \rangle\}$$

This geometry is illustrated in Figure 2.

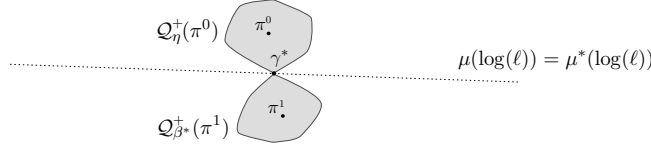


Fig. 2: The Neyman-Pearson hypothesis testing problem. The likelihood ratio test is interpreted as a separating set between the convex sets  $\mathcal{Q}_\eta(\pi^0)$  and  $\mathcal{Q}_{\beta^*}(\pi^1)$ .

**THEOREM 2.1.** *Suppose that  $\{\pi^0, \pi^1\}$  have strictly positive densities on  $\mathbf{X} = \mathbb{R}^d$ , denoted  $\{p^0, p^1\}$ , and suppose that the optimal value of  $I_\phi$  in (2.2) is finite and non-zero. Then the following statements hold,*

- (i) *The optimal value of (2.2) is given by the minimal Kullback-Leibler divergence  $\beta^*$  given in (2.5).*
- (ii) *There exists  $\rho^* > 0$  such that the following alignment condition holds for the optimizer  $\gamma^*$  of (2.4):*

$$\log \frac{d\gamma^*}{d\pi^1}(x) + \rho^* \log \frac{d\gamma^*}{d\pi^0}(x) \leq \beta^* + \rho^* \eta, \quad x \in \mathbf{X},$$

*with equality almost everywhere. Consequently, the optimizer  $\gamma^* \in \mathcal{Q}_\eta^+(\pi^0)$  has density,*

$$q^*(x) = k_0 [p^0(x)]^{\frac{\rho^*}{1+\rho^*}} [p^1(x)]^{\frac{1}{1+\rho^*}}, \quad x \in \mathbf{X}, \quad (2.6)$$

*where  $k_0 > 0$  is a normalizing constant.*

- (iii)  $\beta^* = \max_{\rho \geq 0} \left\{ -\rho \eta - (1 + \rho) \log \left( \int (p^0(x))^{\frac{\rho}{1+\rho}} (p^1(x))^{\frac{1}{1+\rho}} dx \right) \right\}$ ,  
*where the maximum is attained at the value of  $\rho^*$  in (ii).*
- (iv) *The following log-likelihood ratio test (LRT) is optimal, described as the general test (2.3) using the set,*

$$\mathcal{A} := \{\gamma \in \mathcal{M} : \langle \gamma, \log \ell \rangle \leq \beta^* - \eta\}, \quad (2.7)$$

*where  $\ell$  denotes the likelihood ratio  $\ell = d\pi^0/d\pi^1$ .*

*Proof.* Part (i) of Theorem 2.1 is due to Hoeffding [23]. This result follows from Sanov's Theorem as described above.

Parts (ii) and (iii) were established in [6]. We sketch a proof here since similar ideas will be used later in the paper.

To prove (ii) we construct a dual for the convex program (2.5). Consider the relaxation in which the constraint  $\gamma \in \mathcal{Q}_\eta^+(\pi^0)$  is relaxed through the introduction of a Lagrange multiplier  $\rho \in \mathbb{R}_+$ . The Lagrangian is denoted,

$$\mathcal{L}(\gamma; \rho) := D(\gamma \|\pi^1) + \rho(D(\gamma \|\pi^0) - \eta), \quad (2.8)$$

and the dual function  $\Psi: \mathbb{R}_+ \rightarrow \mathbb{R}$  is defined as the infimum of the Lagrangian over all probability distributions,

$$\Psi(\rho) := \inf\{D(\gamma \|\pi^1) + \rho(D(\gamma \|\pi^0) - \eta)\}. \quad (2.9)$$

An optimizer  $\gamma^\rho$  can be characterized by taking directional derivatives. Let  $\gamma \in \mathcal{M}$  be arbitrary, and define  $\gamma^\theta = \theta\gamma + (1-\theta)\gamma^\rho$ . Then the first order condition for optimality is,

$$0 \leq \frac{d}{d\theta} \mathcal{L}(\gamma^\theta; \rho) \Big|_{\theta=0} = \langle \gamma - \gamma^\rho, \log \frac{d\gamma^\rho}{d\pi^1} + \rho(\log \frac{d\gamma^\rho}{d\pi^0} - \eta) \rangle.$$

On setting  $\gamma = \delta_x$ , the point-mass at  $x \in \mathsf{X}$ , we obtain the bound,

$$\log \frac{d\gamma^\rho}{d\pi^1}(x) + \rho \log \frac{d\gamma^\rho}{d\pi^0}(x) \leq D(\gamma^\rho \|\pi^1) + \rho D(\gamma^\rho \|\pi^0),$$

and on integrating both sides of this inequality with respect to  $\gamma^\rho$  we conclude that equality holds a.e.  $[\gamma^\rho]$ . This proves (ii), with  $\gamma^* = \gamma^{\rho^*}$ , where  $\rho^*$  is chosen so that  $\gamma^{\rho^*} \in \mathcal{Q}_\eta^+(\pi^0)$ .

Substituting the optimizer  $\gamma^\rho$  into the Lagrangian (2.8) leads to an explicit expression for the dual,

$$\begin{aligned} \Psi(\rho) &= D(\gamma^* \|\pi^1) + \rho(D(\gamma^* \|\pi^0) - \eta) \\ &= -\eta\rho + \langle \gamma^*, \log \left[ \left( \frac{d\gamma^*}{d\pi^1} \right) \left( \frac{d\gamma^*}{d\pi^0} \right)^\rho \right] \rangle \end{aligned} \quad (2.10)$$

The term within the brackets is constant, which gives,

$$\Psi(\rho) = -\eta\rho - (1 + \rho) \log \left[ \int (p^0(x))^{\frac{\rho}{1+\rho}} (p^1(x))^{\frac{1}{1+\rho}} dx \right].$$

To complete the proof of (iii) we argue that  $\beta^*$  is the value of the dual,

$$\beta^* = \max\{\Psi(\rho) : \rho \geq 0\}.$$

Given this combined with the expression (2.10) for  $\Psi(\rho)$  we obtain the representation (iii).

Part (iv) follows from Sanov's Theorem and the geometry illustrated in Figure 2: The likelihood ratio test is interpreted as a separating set between the convex sets  $\mathcal{Q}_\eta(\pi^0)$  and  $\mathcal{Q}_{\beta^*}(\pi^1)$ .  $\square$

**Robust hypothesis testing** In typical applications it is unrealistic to assume precise values are known for the two marginals  $\pi^0, \pi^1$ . Consider the following relaxation in which hypothesis  $H_i$  corresponds to the assumption that the marginal distribution lies in a closed, affine subset  $\mathbb{P}_i \subset \mathcal{M}$ . A robust N-P hypothesis testing problem is formulated in which the worst-case type-II exponent is maximized over  $\pi_1 \in \mathbb{P}_1$ , subject to a uniform constraint on the type-I exponent over all  $\pi_0 \in \mathbb{P}_0$ :

$$\sup_{\phi} \inf_{\pi_1 \in \mathbb{P}_1} I_{\phi}^{\pi_1} \quad \text{subject to} \quad \inf_{\pi_0 \in \mathbb{P}_0} J_{\phi}^{\pi_0} \geq \eta. \quad (2.11)$$

A test is called optimal if it solves this optimization problem.

The optimization problem (2.11) is considered in [34, 35, 33] in the special case in which the uncertainty sets are defined by specifying a finite number of generalized *moments*: A finite set of real-valued continuous functions  $\{f_j : j = 1, \dots, n\}$  and real constants  $\{c_j^i : i = 0, 1, \dots, n\}$  are given, and

$$\mathbb{P}_i := \{\pi \in \mathcal{M} : \langle \pi, f_j \rangle = c_j^i, \quad j = 0, \dots, n, \quad i = 0, 1\}. \quad (2.12)$$

As a notational convenience we take  $f_0 \equiv 1$  and  $c_0^1 = 1$ .

It is possible to construct a simple optimal test based on a linear function of the data. Although the test itself is not a log-likelihood test, it has a geometric interpretation that is entirely analogous to that given in Theorem 2.1. The value  $\beta^*$  in an optimal test can be expressed,

$$\beta^* = \inf\{\beta : \mathcal{Q}_{\eta}^+(\mathbb{P}_0) \cap \mathcal{Q}_{\beta}^+(\mathbb{P}_1) \neq \emptyset\}. \quad (2.13)$$

Moreover, the infimum is achieved by some  $\mu^* \in \mathcal{Q}_{\eta}^+(\mathbb{P}_0) \cap \mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$ , along with *least favorable* distributions  $\pi_0^* \in \mathbb{P}_0, \pi_1^* \in \mathbb{P}_1$ , satisfying

$$D(\mu^* \parallel \pi_0^*) = \eta, \quad D(\mu^* \parallel \pi_1^*) = \beta^*.$$

The distribution  $\mu^*$  has the form  $\mu^*(x) = \ell_0(x)\pi_0^*(x)$ , where the function  $\ell_0$  is a linear combination of the constraint functions  $\{f_i\}$  used to define  $\mathbb{P}_0$ . The function  $\log \ell_0$  defines a separating hyperplane between the convex sets  $\mathcal{Q}_{\eta}^+(\mathbb{P}_0)$  and  $\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$ , as illustrated in Figure 3.

Note that  $\log \ell_0$  is defined everywhere, yet in applications the likelihood ratio  $d\mu^*/d\pi_0^*$  may be defined only on a small subset of  $\mathcal{X}$ .

**PROPOSITION 2.1.** *Suppose that the moment classes  $\mathbb{P}_0$  and  $\mathbb{P}_1$  each satisfy the non-degeneracy condition that the vector  $(c_0^i, \dots, c_n^i)$  lies in the relative interior of the set of all possible moments  $\{\mu(f_0, \dots, f_n) : \mu \in \mathcal{M}\}$ . Then, there exists  $\{\lambda_0, \dots, \lambda_n\} \in \mathbb{R}$  such that the function  $\ell_0 = \sum \lambda_i f_i$  is non-negative valued, and the following test is optimal*

$$\phi_N^* = 0 \iff \frac{1}{N} \sum_{t=0}^{N-1} \log(\ell_0(X_t)) \leq \eta. \quad (2.14)$$

*Proof.* The result is given as [34, Proposition 2.4]. We sketch the proof here, based on the convex geometry illustrated in Figure 3.

Since  $\mathcal{Q}_\eta^+(\mathbb{P}_0)$  and  $\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$  are compact sets it follows from their construction that there exists  $\mu^* \in \mathcal{Q}_\eta^+(\mathbb{P}_0) \cap \mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$ . Moreover, by convexity there exists *some* function  $h: \mathbf{X} \rightarrow \mathbb{R}$  defining a separating hyperplane between the sets  $\mathcal{Q}_\eta^+(\mathbb{P}_0)$  and  $\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$ , satisfying

$$\mathcal{Q}_\eta(\mathbb{P}_0) \subset \{\mu \in \mathcal{M} : \langle \mu, h \rangle < \eta\}, \quad \mathcal{Q}_{\beta^*}(\mathbb{P}_1) \subset \{\mu \in \mathcal{M} : \langle \mu, h \rangle > \eta\}.$$

The remainder of the proof consists of the identification of  $h$  using the Kuhn-Tucker alignment conditions based on consideration of a dual functional as in the proof of Theorem 2.1.  $\square$

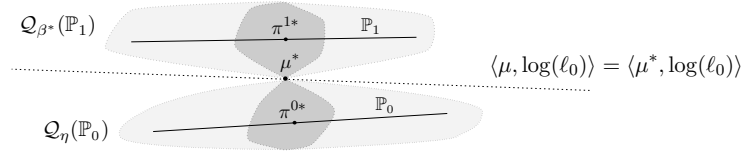


Fig. 3: The two-moment worst-case hypothesis testing problem. The uncertainty classes  $\mathbb{P}_i$ ,  $i = 0, 1$  are determined by a finite number of linear constraints, and the thickened regions  $\mathcal{Q}_\eta(\mathbb{P}_0)$ ,  $\mathcal{Q}_{\beta^*}(\mathbb{P}_1)$  are each convex. The linear threshold test is interpreted as a separating hyperplane between these two convex sets.

**2.2. Mutual information.** In this section we derive the expression (1.5) for channel capacity based on Theorem 2.1, following ideas in Anantharam [3] (see also [12, 15].)

Consider the decoding problem in which a set of  $N$ -dimensional code words are generated by a sequence of random variables with marginal distribution  $\mu$ . The receiver is given the output sequence  $\{Y_1, \dots, Y_N\}$  and considers an arbitrary sequence from the code book  $\{X_1^i, \dots, X_N^i\}$ , where  $i$  is the index in a finite set  $\{1, \dots, e^{NR}\}$ , where  $R$  is the rate of the code. Since  $\mathbf{X}^i$  has marginal distribution  $\mu$ ,  $\mathbf{Y}$  has marginal distribution  $p_\mu$  defined in (1.2).

For each  $i$ , this decision process can be interpreted as a binary hypothesis testing problem in which  $H_0$  is the hypothesis that  $\{(X_j^i, Y_j) : j = 1, \dots, N\}$  has marginal distribution

$$\pi^0 := \mu \otimes p_\mu.$$

That is,  $\mathbf{X}^i$  and  $\mathbf{Y}$  are independent if codeword  $i$  was not sent. Hypothesis  $H_1$  is the hypothesis that  $i$  is the true code word, so that the joint marginal distribution is

$$\pi^1 := \mu \odot p[dx, dy] := \mu(dx)p(y|x)dy.$$

Suppose that the error exponent  $\eta > 0$  is given, and an optimal N-P LRT test is applied. Then  $I_\phi = \eta$  means that,

$$\begin{aligned} \eta &= - \lim_{N \rightarrow \infty} \frac{1}{N} \log(\mathbb{P}_{\pi^0}(\phi_N(X_1, \dots, X_N) = 1)) \\ &= - \lim_{N \rightarrow \infty} \frac{1}{N} \log(\mathbb{P}\{\text{Code word } i \text{ is accepted} \mid i \neq i^*\}), \end{aligned}$$

where the index  $i^*$  denotes the code word sent.

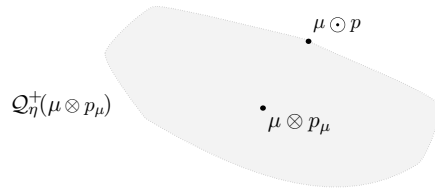


Fig. 4: The channel capacity is equal to the maximal relative entropy between  $p_\mu \otimes \mu$  and  $p_\mu \odot P$ , over all input distributions  $\mu$  satisfying the given constraints.

Consideration of  $e^{RN}$  codewords, our interest lies in the probability that at least one of the  $e^{RN} - 1$  incorrect code words is mistaken for the true code word. We obtain through the union bound,

$$\begin{aligned} &\mathbb{P}\{\text{The true code word } i^* \text{ is rejected}\} \\ &\leq \sum_{i \neq i^*} \lim_{N \rightarrow \infty} \mathbb{P}\{\text{Code word } i \text{ is accepted} \mid i \neq i^*\}, \end{aligned}$$

from which we obtain,

$$\eta - R \geq - \lim_{N \rightarrow \infty} \frac{1}{N} \log(\mathbb{P}\{\text{The true code word } i^* \text{ is rejected}\}). \quad (2.15)$$

We must have  $R < \eta$  to ensure that right hand side is positive, so that the probability that the true code word  $i^*$  is rejected vanishes as  $N \rightarrow \infty$ .

One must also ensure that  $\eta$  is not too large, since it is necessary that  $\beta^* > 0$  so that  $J_\phi > 0$  under the LRT. Hence an upper bound on  $R$  is the supremum over  $\eta$  satisfying  $\beta^* > 0$ , which is precisely mutual information:

$$I(\mu) := D(\mu \odot P \parallel \mu \otimes p_\mu) = \int \int \log\left(\frac{p(y|x)dy}{p_\mu(dy)}\right) \mu(dx)p(y|x)dy. \quad (2.16)$$

This conclusion is illustrated in Figure 4.

The channel capacity is defined to be the maximum of  $I$  over all input distributions  $\mu$  satisfying the given constraints. We thus arrive at the convex program (1.5).

**2.3. Error exponents.** A representation of the channel-coding random coding exponent can be obtained based on similar reasoning. Here we illustrate the form of the solution, and show that it may be cast as a *robust* hypothesis testing problem of the form considered in Section 2.1.

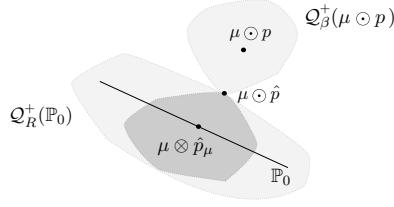


Fig. 5: The error exponent is equal to the solution of a robust N-P hypothesis testing problem.

For a given  $\mu \in \mathcal{M}$ , denote by  $\mathbb{P}_0$  the space of product measures on  $\mathsf{X} \times \mathsf{Y}$ ,

$$\mathbb{P}_0 = \{\mu \otimes \nu : \nu \text{ is a probability measure on } \mathsf{Y}\},$$

and define the corresponding divergence set for a given  $R > 0$ ,

$$\mathcal{Q}_R^+(\mathbb{P}_0) := \bigcup_{\nu} \mathcal{Q}_R^+(\mu \otimes \nu).$$

Equivalently,  $\mathcal{Q}_R^+(\mathbb{P}_0) = \{\gamma : \min_{\nu} D(\gamma \parallel \mu \otimes \nu) \leq R\}$ . The robust hypothesis testing problem is binary, with  $H_1$  as defined in the channel capacity problem, but with  $H_0$  defined using  $\mathbb{P}_0$ :

$H_0$ :  $\{(X_j^i, Y_j) : j = 1, \dots, N\}$  has marginal distribution  $\pi^0 \in \mathbb{P}_0$ .

$H_1$ :  $\{(X_j^i, Y_j) : j = 1, \dots, N\}$  has marginal distribution  $\pi^1 := \mu \circledast p$ .

Proposition 2.2 shows that the random coding exponent  $E_r(R)$  can be represented as the solution to the robust N-P hypothesis testing problem (2.11) with  $\eta = R$  and  $\mathbb{P}_1 = \{\mu \circledast p\}$ .

PROPOSITION 2.2.

$$E_r(R) = \sup_{\mu} \left( \inf_{\beta} \{\beta : \mathcal{Q}_{\beta}^+(\mu \circledast p) \cap \mathcal{Q}_R^+(\mathbb{P}_0) \neq \emptyset\} \right). \quad (2.17)$$

Suppose that there exists a triple  $(\mu^*, \nu^*, \gamma^*)$  that solve (2.17) in the sense that

$$D(\gamma^* \parallel \mu^* \circledast p) = E_r(R), \quad D(\gamma^* \parallel \mu^* \otimes \nu^*) = R.$$

Then, there exists a channel transition density  $\hat{p}$  such that

$$\gamma^* = \mu^* \circledast \hat{p}, \quad \nu^* = \hat{p}_{\mu^*},$$

and the rate can be expressed as mutual information,

$$R = I(\mu^*; \hat{p}) := D(\mu^* \odot \hat{p} \parallel \mu^* \otimes \hat{p}_\mu).$$

*Proof.* Blahut in [6] establishes several representations for  $E_r$ , beginning with the following

$$E_r(R) = \sup_{\mu} \inf_{\mu \odot \hat{p} \in \hat{\mathcal{Q}}_R^+} D(\mu \odot \hat{p} \parallel \mu \odot p) \quad (2.18)$$

where the supremum is over all  $\mu$ , subject to the given constraints, and the infimum is over all transition densities  $\hat{p}$  satisfying  $\mu \odot \hat{p} \in \hat{\mathcal{Q}}_R^+$  where

$$\hat{\mathcal{Q}}_R^+ := \{\mu \odot \hat{p} : D(\mu \odot \hat{p} \parallel \mu \otimes \hat{p}_\mu) \leq R\}.$$

The optimization problem (2.17) is a relaxation of (2.18) in which the distributions  $\{\nu\}$  in the definition of  $\mathbb{P}_0$  are constrained to be of the form  $\hat{p}_\mu$ , and the distributions  $\{\gamma\}$  are constrained to be of the form  $\mu \odot \hat{p}$  for some transition density  $\hat{p}$ .

It remains to show that these restrictions hold for any solution to (2.17), so that the relaxed optimization problem (2.17) is equivalent to (2.18).

For fixed  $\mu$ , denote the infimum over  $\beta$  in (2.17) by,

$$\beta^*(\mu) := \inf\{\beta : \mathcal{Q}_\beta^+(\mu \odot p) \cap \mathcal{Q}_R^+(\mathbb{P}_0) \neq \emptyset\} \quad (2.19)$$

If  $(\nu^*, \gamma^*)$  solve (2.19) in the sense that

$$D(\gamma^* \parallel \mu \odot p) = \beta^*(\mu), \quad D(\gamma^* \parallel \mu \otimes \nu^*) = R,$$

then the distribution  $\gamma^*$  solves the ordinary N-P hypothesis testing problem with  $\pi^0 = \mu \otimes \nu^*$  and  $\pi^1 = \mu \odot p$ . It then follows that the first marginal  $\gamma_1^*$  is equal to  $\mu$  by the representation given in Theorem 2.1 (ii).

Moreover, the second marginal of  $\gamma_2^*$  can be taken to be  $\nu^*$  since for any  $\nu$ ,

$$D(\gamma^* \parallel \mu \otimes \nu) = D(\gamma^* \parallel \mu \otimes \gamma_2^*) + D(\gamma_2^* \parallel \nu) \geq D(\gamma^* \parallel \mu \otimes \gamma_2^*).$$

These conclusions imply that  $\gamma^* = \mu \odot \hat{p}$  for some channel density  $\hat{p}$ , and  $\nu^* = \hat{p}_\mu$ , which establishes the desired equivalence between the optimization problems (2.17) and (2.18).  $\square$

The solution to the optimization problem (2.19) is illustrated in Figure 5. The channel transition density  $\hat{p}$  shown in the figure solves

$$\begin{aligned} \beta^*(\mu) &= \inf\{\beta : \mathcal{Q}_\beta^+(\mu \odot p) \cap \mathcal{Q}_R^+(\mu \otimes \hat{p}_\mu) \neq \emptyset\} \\ &= D(\mu \odot \hat{p} \parallel \mu \odot p). \end{aligned}$$

The error exponent is equal to the maximal relative entropy  $\beta^*(\mu)$  over all  $\mu$ , and the rate can be expressed as mutual information  $R = I(\mu^*; \hat{p}) := D(\mu^* \odot \hat{p} \parallel \mu^* \otimes \hat{p}_\mu)$  where  $\mu^*$  is the optimizing distribution.

**3. Convex optimization and channel coding.** The alignment conditions for the N-P hypothesis testing problem were derived in Theorem 2.1 based on elementary calculus. Similar reasoning leads to alignment conditions characterizing channel capacity and the optimal error exponent.

**3.1. Mutual information.** Recall that the channel sensitivity function  $g_\mu(x)$  is the point-wise relative entropy,  $g_\mu(x) := D(p(\cdot | x) \| p(\cdot | \mu))$ . The next result, taken from [25], shows that  $g_\mu$  is the gradient of  $I$  at  $\mu$ .

PROPOSITION 3.1. *For any given  $\mu, \mu^\circ \in \mathcal{M}(\sigma_P^2, M, \mathbf{X})$ ,*

- (i)  $I(\mu) = \langle \mu, g_\mu \rangle = \max_{\mu' \in \mathcal{M}} \langle \mu, g_{\mu'} \rangle$ .
- (ii) *Letting  $\mu_\theta := (1 - \theta)\mu^\circ + \theta\mu$ , the first-order sensitivity with respect to  $\theta \in [0, 1]$  is*

$$\left. \frac{d}{d\theta} I(\mu_\theta) \right|_{\theta=0} = \langle \mu - \mu^\circ, g_{\mu^\circ} \rangle \quad (3.1)$$

□

The dual functional  $\Psi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is defined by

$$\Psi(r) = \sup_{\mu \in \mathcal{M}_0} [I(\mu) - r\langle \mu, \phi \rangle], \quad r \geq 0, \quad (3.2)$$

where  $\mathcal{M}_0 = \mathcal{M}(M^2, M, \mathbf{X}) = \mathcal{M}(\infty, M, \mathbf{X})$  denotes the constraint set without an average power constraint. The dual functional is a convex, decreasing function of  $r$ , as illustrated in Figure 6. Note that we do not exclude  $M = \infty$ . In this case,  $\mathcal{M}_0 = \mathcal{M}$ , which denotes the set of probability distributions on  $\mathbf{X}$ .

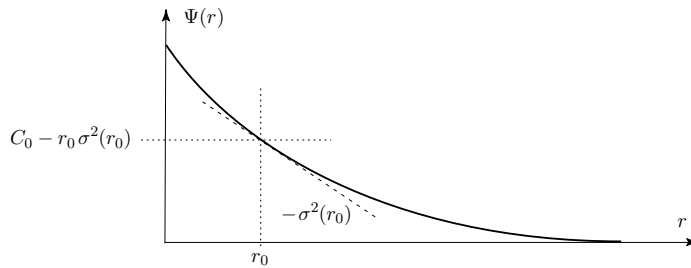


Fig. 6: The dual functional is convex and decreasing. For a given  $r_0 > 0$ , the slope determines an average power constraint  $\sigma^2(r_0)$ , and the corresponding capacity  $C_0 := C(\sigma^2(r_0), M, \mathbf{X})$  may be determined as shown in the figure.

The parameter  $r$  provides a convenient parameterization of the optimization problem (1.5). The proof of Theorem 3.1. may be found in [25, Theorem 2.8].

THEOREM 3.1. *If  $M < \infty$ , then an optimizing distribution  $\mu_r^*$  exists for (3.2) for each  $r > 0$ , and the following hold:*

(i) *The alignment condition holds,*

$$g_{\mu_r^*}(x) \leq \Psi(r) + rx^2, \quad |x| \leq M,$$

*with equality a.e.  $[\mu_r^*]$ .*

(ii) *Let  $\sigma^2(r) := -\frac{d}{dr}\Psi(r)$ . The distribution  $\mu_r^*$  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*

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

□

**3.2. Error exponents.** Boundedness of the sensitivity function is central to the analysis of [25].

LEMMA 3.1.  $0 < g_\mu^\rho(x) \leq 1$  for each  $x$ , and  $g_\mu^\rho \rightarrow 0$  as  $x \rightarrow \infty$ . □

Continuity of  $G^\rho$  follows easily from Lemma 3.1. The following set of results establishes convexity and differentiability of  $G^\rho$  with respect to  $\mu$ . For  $\mu, \mu^\circ \in \mathcal{M}$  and  $\theta \in [0, 1]$  we denote  $\mu_\theta := (1 - \theta)\mu^\circ + \theta\mu$ .

PROPOSITION 3.2. *For any given  $\mu, \mu^\circ \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$  and  $\rho > 0$ ,*

- (i) *For a given  $\rho$ , the mapping  $G^\rho: \mathcal{M}(\sigma_P^2, M, \mathsf{X}) \mapsto \mathbb{R}_+$  is continuous in the weak topology.*
- (ii) *The functional  $G^\rho$  is convex, and can be expressed as the maximum of linear functionals,*

$$G^\rho(\mu^\circ) = \langle \mu, g_\mu^\rho \rangle = \max_{\mu \in \mathcal{M}} \{(1 + \rho)\langle \mu^\circ, g_\mu^\rho \rangle - \rho G^\rho(\mu)\}$$

(iii) *Fix  $\rho \geq 0$ ,  $\mu^\circ \in \mathcal{M}$ . The first order sensitivity is given by*

$$\left. \frac{d}{d\theta} G^\rho(\mu_\theta) \right|_{\theta=0} = (1 + \rho)\langle \mu - \mu^\circ, g_{\mu^\circ}^\rho \rangle.$$

□

For fixed  $\rho$ , the optimization problem (1.9) is a convex program since  $G^\rho$  is convex. Continuity then leads to existence of an optimizer. The following result from [26] summarizes the structure of the optimal input distribution. It is similar to Theorem 3.1 which requires the peak power constraint  $M < \infty$ . We stress that this condition is not required here.

THEOREM 3.2. *For each  $\rho \geq 0$ , there exists  $\mu^\circ \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$  that achieves  $G^{\rho*}$ .*

*Moreover, a distribution  $\mu^\circ \in \mathcal{M}(\sigma_P^2, M, \mathsf{X})$  is optimal if and only if there exists a real number  $\lambda_1^*$  and a positive real number  $\lambda_2^*$  such that*

$$g_{\mu^\circ}^\rho(x) \geq \lambda_1^* - \lambda_2^*x^2, \quad x \in \mathsf{X}, \quad \text{and} \quad g_{\mu^\circ}^\rho(x) = \lambda_1^* - \lambda_2^*x^2, \quad \text{a.e. } [\mu^\circ]$$

If these conditions hold, then

$$G^{\rho*} := \min_{\mu} G^{\rho}(\mu) = G^{\rho}(\mu^{\circ}) = \frac{\lambda_1^* - \lambda_2^* \sigma_P^2}{1 + \rho}.$$

□

Shown in Figure 7 are plots of  $g_{\mu}^{\rho}$  for two distributions. The input distribution  $\mu_0$  violates the alignment condition in Theorem 3.2 (ii), and hence is not optimal. The alignment condition does hold for  $\mu_1$ , and we conclude that this distribution does optimize (1.5).

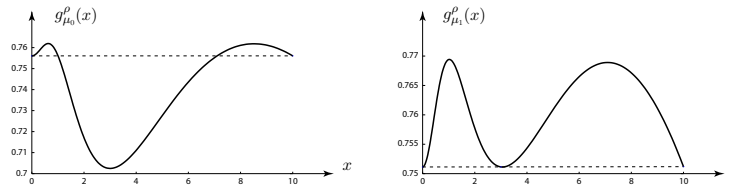


Fig. 7: Numerical results for the Rayleigh channel  $Y = AX + N$  subject to peak-power constraint for rate  $R < C$ , with  $\sigma_A^2 = \sigma_N^2 = 1$  and  $\rho = 0.5$ .

**PROPOSITION 3.3.** *For given  $\rho$ , any optimal input distribution  $\mu^*$  achieving  $G^{\rho*}$  is discrete, with a finite number of mass points in any interval.*

*Proof.* To see that the optimizer  $g_{\mu^*}^{\rho}$  is discrete consider the alignment conditions. There exists a quadratic function  $q^*$  satisfying  $q^*(x) \leq g_{\mu^*}^{\rho}(x)$ , with equality a.e.  $[\mu^*]$ . Lemma 3.1 asserts that  $g_{\mu^*}^{\rho}$  takes on strictly positive values and vanishes at infinity. It follows that  $q^*$  is not a constant function, and hence  $q^*(x) \rightarrow -\infty$  as  $x \rightarrow \infty$ . This shows that the optimizer has bounded support, with

$$\text{supp}(\mu^*) \subset \{x : q^*(x) > 0\} = \{x : x^2 \leq \sigma_P^2 + \lambda_0/\lambda_2\}.$$

Moreover, since  $g_{\mu^*}^{\rho}$  is an analytic function on  $\mathbb{X}$  it then follows that  $g_{\mu^*}^{\rho}(x) = q^*(x)$  is only possible for a finite number of points. □

**Optimal binary distributions** Gallager in [19] bounds the random coding exponent by a linear functional over the space of probability measures. The bound is shown to be tight for low SNR, and thus the error exponent optimization problem is converted to a linear program over the space of probability measures. An optimizer is an extreme point, which is shown to be binary. Similar arguments used in [25] can be generalized to the model considered here.

We begin with consideration of *zero* SNR, which leads us to consider the sensitivity function using  $\mu = \delta_0$ , the point mass at 0, denoted  $g_0^{\rho} := g_{\delta_0}^{\rho}(x)$ . It is easy to see  $g_0^{\rho}(0) = 1$ , and we have seen that  $g_0^{\rho}(x) \leq 1$  everywhere. Given the analyticity assumption, it follows that this function

has zero derivative at the origin, and non-positive second derivative. We thus obtain the bound,

$$\frac{d \log(1 - g_0^\rho(x))}{d \log(x)} \Big|_{x=0} \geq 2,$$

with equality holding if and only if the second derivative of  $g_0^\rho(x)$  is non-zero at  $x = 0$ .

We have the following proposition concerning the binary distribution at low SNR. This extends Theorem 3.4 of [25] which covers channel capacity. However, unlike this previous result and the result of [19], we do not require a peak power constraint on the input distribution.

**PROPOSITION 3.4.** *Consider a channel with  $X = \mathbb{R}_+$ . For a fixed  $\rho > 0$ , suppose that the following hold,*

- (i)  $\frac{d^2}{dx^2} g_0^\rho(0) = 0$ .
- (ii) *There is a unique  $x_1 > 0$  satisfying*

$$\frac{d \log(1 - g_0^\rho(x))}{d \log(x)} \Big|_{x=x_1} = 2.$$

- (iii) *There is ‘positive sensitivity’ at  $x_1$ :*

$$\frac{d}{dx} \left( \frac{d \log(1 - g_0^\rho(x))}{d \log(x)} \right) \Big|_{x=x_1} \neq 0.$$

*Then, for all SNR sufficiently small, the optimal input distribution is binary with one point at the origin.  $\square$*

The proof of Proposition 3.4 may be found in [26]. We illustrate the proof and its assumptions using the Rayleigh channel.

Given the channel transition probability function (1.15), the sensitivity function is,

$$g_0^\rho(x) = \frac{(1 + \rho)(1 + x^2)^{\frac{\rho}{1+\rho}}}{(1 + x^2)\rho + 1}, \quad x \geq 0.$$

From the plot shown at left in Figure 8 we see that there exists a quadratic function  $q_0$  satisfying  $q_0(x) \leq g_0^\rho(x)$ , with equality at the origin and precisely one  $x_1 > 0$ . The plot at right shows that (iii) holds, and hence that all of the assumptions of Proposition 3.4 are satisfied. Consequently, with vanishing SNR, the optimizing distribution is binary, and approximates the binary distribution supported on  $\{0, x_1\}$ .

**4. Signal constellation design.** We now show how the conclusions of this paper may be applied in design.

For a symmetric, complex channel we have seen in Theorem 1.1 that the optimal input distribution is circularly symmetric on  $\mathbb{C}$ , and discrete

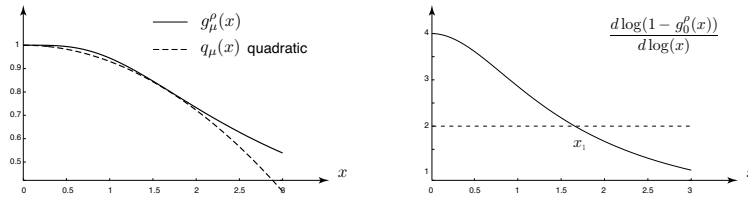


Fig. 8: Optimal binary distribution for the Rayleigh channel. At left is a plot of  $g_0^\rho$  together with the quadratic function aligned with  $g_0^\rho$ . The two functions meet at only two points. Shown at right is a plot of the log-derivative: The nonzero point of support for the optimal binary input is found by setting this equal to 2.

in magnitude. We consider in this section discrete approximations of this optimal distribution on  $\mathbb{C}$ .

We propose the following approach to signal constellation design and coding in which the signal alphabet and associated probabilities are chosen to provide a random code that approximates the random code obtained through the nonlinear program (1.5). We conclude this paper with examples to illustrate the performance of this approach, as compared to standard approaches based on QAM or PSK.

**Complex AWGN channel** This is the complex channel model given by  $Y = X + N$  with  $N$  complex Gaussian. Examples considered in [7] suggest that QAM typically outperforms PSK when the constellation sizes are fixed, and the signal to noise ratio is large. For small SNR, it is known that QAM and PSK are almost optimal (see [7, Figure 7.11], [40], and related results in [37]).

Figure 9 shows results using two signal constellation methods: 4-point QAM and a 5-point distribution which contains the 4-point QAM plus a point at origin. The 5 point distribution is an approximation to the optimal input distribution, which is binary in magnitude, and uniformly distributed in phase. The 5-point constellation performs better than 4-point QAM by about 13%, with lower power consumption.  $\square$

**Rayleigh channel with low SNR** We consider the normalized model in which  $A, N$  are each Gaussian, mutually independent, and circularly symmetric, with  $\sigma_A^2 = 1, \sigma_N^2 = 1$ . Consideration of the magnitude of  $X$  and  $Y$  leads to the real channel model with transition density shown in (1.15).

We compare codes obtained from the two constellations illustrated in Figure 10. The first constellation is a 16-point QAM. Since the code used in QAM is a random code with uniform distribution, the average power is given by  $\sigma_P^2 = 11.7$ . The second constellation has only two elements: one point at origin and another point at position  $5 \in \mathbb{C}$ . The weights are chosen so that the average power is again  $\sigma_P^2 = 11.7$ , which results in

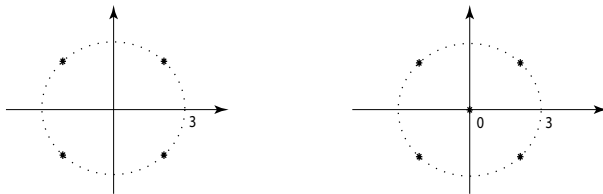


Fig. 9: Left: the 4-point QAM signal constellation for complex AWGN channel  $Y = X + N$ , with  $\sigma_P^2 = 9$  and  $\sigma_N^2 = 1$ , i.e.  $\text{SNR} = 9$  (9.54dB). The mutual information achieved by this 4-point QAM is 1.38 nats/symbol. Right: A 5-point constellation signal constellation for complex AWGN channel  $Y = X + N$ , with  $\sigma_N^2 = 1$ , with 4 points (with equal probability) at the same position as QAM plus one point at origin with probability equal to 0.1077. The constellation achieves 1.52 nats/symbol mutual information.

$\mu\{0\} = 1 - \mu\{5\} = 0.5346$ . This is the optimal input distribution when the peak-power constraint  $M = 5$  is imposed.

Computations show that the simpler coding scheme achieves mutual information 0.4879 nats/symbol, which is about 2.5 times more than the mutual information achieved by the 16-point QAM code.  $\square$

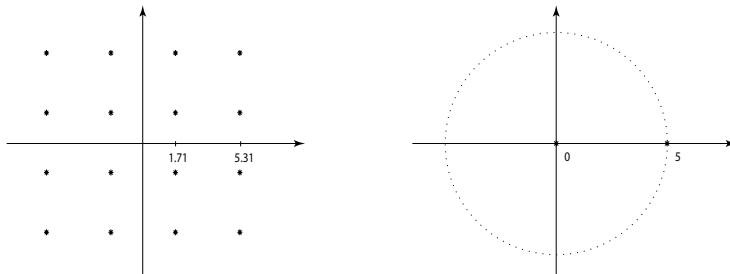


Fig. 10: Left: the 16-point QAM signal constellation for Rayleigh channel  $Y = AX + N$ , with  $\sigma_A^2 = 1$ ,  $\sigma_N^2 = 1$ , and average power constraint  $\sigma_P^2 = 11.7$ . The mutual information achieved is 0.1951 nats/symbol. Right: A 2-point constellation with one point at origin (with probability 0.5346) and another point with magnitude 5, for the same channel model and average power constraint. The mutual information achieved is 0.4879 nats/symbol, which is 2.5 times more than that achieved by the 16-point QAM.

**Rayleigh channel with high SNR** In this final example the same parameters used in the previous experiment are maintained, except now the average power is increased to  $\sigma_P^2 = 26.4$ . The optimal input distribution is given as follows when the channel is subject to the peak power constraint  $|X| \leq 8$ : The phase may be taken uniformly distributed without any loss of generality, and the magnitude has three points of support at  $\{0.0, 2.7, 8.0\}$  with respective probabilities  $\{0.465, 0.138, 0.397\}$ . Consequently, we propose a constellation whose magnitude is restricted to these three radii. This is compared to 16-point QAM. The two constellation designs are illustrated

in Figure 11.

If the probability weights  $\{0.465, 0.138, 0.397\}$  are used in the proposed constellation design, then the resulting mutual information is 0.5956 nats/symbol, which is about 3 times larger than the mutual information achieved by the 16-point QAM.  $\square$

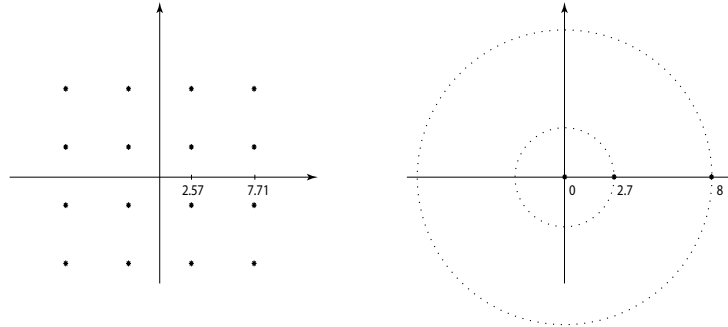


Fig. 11: The plot at left shows the 16-point QAM signal constellation, and at right is shown a three-point constellation with one point at origin; one point on a circle of radius 2.57; and the third point on a circle of radius 8. The respective probabilities are uniform for the QAM code, and given by  $(0.5346, 0.1379, 0.397)$  for the respective codewords in the three-point constellation.

**5. Conclusions.** Many problems in information theory may be cast as a convex program over a set of probability distributions. Here we have seen three: hypothesis testing, channel capacity, and computation of the random coding exponent. Another example considered in [24] is computation of the distortion function in source coding. Although the optimization problem in each case is infinite dimensional when the state space is not finite, in each example we have considered it is possible to construct a finite dimensional algorithm, and convergence is typically very fast. We believe this is in part due to the extremal nature of optimizers. Since optimizers have few points of support, this means the optimizer is on the boundary of the constraint set, and hence sensitivity is typically non-zero.

There are many unanswered questions:

- (i) The theory described here sets the stage for further research on channel sensitivity. For example, how sensitive is the error exponent to SNR, coherence, channel memory, or other parameters.
- (ii) It is possible to extend most of these results to multiple access channels. However, we have not yet extended the cutting plane algorithm to MIMO channels, and we don't know if the resulting algorithms will be computationally feasible.
- (iii) Can we apply the results and algorithms we have here to adaptively construct efficient constellations for fading channels?

## REFERENCES

- [1] I. C. Abou-Faycal, M. D. Trott, and S. Shamai. The capacity of discrete-time memoryless Rayleigh-fading channels. *TIT*, 47(4):1290–1301, May 2001.
- [2] 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.
- [3] V. Anantharam. A large deviations approach to error exponents in source coding and hypothesis testing. *IEEE Trans. Inform. Theory*, 36(4):938–943, 1990.
- [4] R.R. Bahadur. *Some Limit Theorems in Statistics*. SIAM, Philadelphia, PA, 1971.
- [5] D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 1999.
- [6] R. E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Information Theory*, IT-20:405–417, 1974.
- [7] R.E Blahut. *Principles and Practice of Information Theory*. McGraw-Hill, New York, 1995.
- [8] J.M. Borwein and A.S. Lewis. A survey of convergence results for maximum entropy. In A. Mohammad-Djafari and G. Demoment, editors, *Maximum Entropy and Bayesian Methods*, pages 39–48. Kluwer Academic, Dordrecht, 1993.
- [9] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
- [10] T.H. Chan, S. Hranilovic, and F.R. Kschischang. Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs. *to appear on IEEE Trans. Inform. Theory*, 2004.
- [11] Rong-Rong Chen, B. Hajek, R. Koetter, and U. Madhow. On fixed input distributions for noncoherent communication over high SNR Rayleigh fading channels. *IEEE Trans. Inform. Theory*, 50(12):3390–3396, 2004.
- [12] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
- [13] I. Csiszár. Sanov property, generalized  $I$ -projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, 1984.
- [14] I. Csiszár. The method of types. *IEEE Trans. Inform. Theory*, 44(6):2505–2523, 1998. Information theory: 1948–1998.
- [15] A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications*. Springer-Verlag, New York, second edition, 1998.
- [16] Paul Dupuis and Richard S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, 1997. A Wiley-Interscience Publication.
- [17] S.-C. Fang, J. R. Rajasekera, and H.-S. J. Tsao. *Entropy optimization and mathematical programming*. International Series in Operations Research & Management Science, 8. Kluwer Academic Publishers, Boston, MA, 1997.
- [18] R.G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.
- [19] R.G. Gallager. Power limited channels: Coding, multiaccess, and spread spectrum. In R.E. Blahut and R. Koetter, editors, *Codes, Graphs, and Systems*, pages 229–257. Kluwer Academic Publishers, Boston, 2002.
- [20] J.D. Gibson, R.L. Baker, T. Berger, T. Lookabaugh, and D. Lindbergh. *Digital Compression for Multimedia*. Morgan Kaufmann Publishers, San Francisco, CA, 1998.
- [21] M.C. Gursoy, H.V. Poor, and S. Verdu. The noncoherent Rician fading channel - part I: Structure of capacity achieving input. *IEEE Trans. Wireless Communication (to appear)*, 2005.
- [22] M.C. Gursoy, H.V. Poor, and S. Verdu. The noncoherent Rician fading channel - part II: Spectral efficiency in the low power regime. *IEEE Trans. Wireless Communication (to appear)*, 2005.
- [23] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *Ann.*

- Math. Statist.*, 36:369–408, 1965.
- [24] J. Huang. *Characterization and computation of optimal distribution for channel coding*. PhD thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois, 2004.
  - [25] J. Huang and S. P. Meyn. Characterization and computation of optimal distribution for channel coding. *IEEE Trans. Inform. Theory*, 51(7):1–16, 2005.
  - [26] J. Huang, S. P. Meyn, and M. Medard. Error exponents for channel coding and signal constellation design. Submitted for publication, October 2005.
  - [27] M. Katz and S. Shamai. On the capacity-achieving distribution of the discrete-time non-coherent additive white gaussian noise channel. In *Proc. IEEE Int'l. Symp. Inform. Theory, Lausanne, Switzerland, June 30 - July 5.*, page 165, 2002.
  - [28] M. Katz and S. Shamai. On the capacity-achieving distribution of the discrete-time non-coherent additive white Gaussian noise channel. In *2002 IEEE International Symposium on Information Theory*, page 165, 2002.
  - [29] S. Kullback. *Information Theory and Statistics*. Dover Publications Inc., Mineola, NY, 1997. Reprint of the second (1968) edition.
  - [30] A. Lapidoth and S.M. Moser. Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels. *IEEE Trans. Inform. Theory*, 49(10), Oct. 2003.
  - [31] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. available from <http://www.inference.phy.cam.ac.uk/mackay/itila/>.
  - [32] R. Palanki. On the capacity-achieving distributions of some fading channels. Presented at 40th Allerton Conference on Communication, Control, and Computing, 2002.
  - [33] C. Pandit. *Robust Statistical Modeling Based On Moment Classes With Applications to Admission Control, Large Deviations and Hypothesis Testing*. PhD thesis, University of Illinois at Urbana Champaign, University of Illinois, Urbana, IL, USA, 2004.
  - [34] C. Pandit and S. P. Meyn. Worst-case large-deviations with application to queueing and information theory. To appear, *Stoch. Proc. Applns.*, 2005.
  - [35] C. Pandit, S. P. Meyn, and V. V. Veeravalli. Asymptotic robust Neyman-Pearson testing based on moment classes. In *Proceedings of the International Symposium on Information Theory (ISIT), 2004*, June 2004.
  - [36] J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.
  - [37] 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.
  - [38] J. G. Smith. The information capacity of amplitude and variance-constrained scalar gaussian channels. *Inform. Contr.*, 18:203–219, 1971.
  - [39] S. Verdú. On channel capacity per unit cost. *IEEE Trans. Inform. Theory*, 36(5):1019–1030, 1990.
  - [40] S. Verdú. Spectral efficiency in the wideband regime. *IEEE Trans. Inform. Theory*, 48(6):1319–1343, June 2002.
  - [41] Ofer Zeitouni and Michael Gutman. On universal hypotheses testing via large deviations. *IEEE Trans. Inform. Theory*, 37(2):285–290, 1991.