

Extremal Distributions in Information Theory and Hypothesis Testing

Charuhas Pandit, Jianyi Huang, Sean Meyn, Venu Veeravalli

Department of Electrical and Computer Engineering
and the Coordinated Sciences Laboratory,
University of Illinois at Urbana-Champaign
charuhas.pandit@morganstanley.com, {jhuang7, meyn, vvv}@uiuc.edu

I. INTRODUCTION

Many problems in Information Theory can be distilled to an optimization problem over a space of probability distributions. The most important examples are in communication theory, where it is necessary to maximize mutual information in order to compute channel capacity, and the classical hypothesis testing problem in which an optimal test is based on the maximization of divergence.

Two general classes of optimization problems are considered in this paper: convex and linear programs, where the constraint set is defined by a finite number of moment constraints. Let $\mathbf{X} \subset \mathbb{R}$ denote a compact set, \mathcal{M} the set of finite distributions on $\mathcal{B}(\mathbf{X})$, and $\mathcal{M}_1 \subset \mathcal{M}$ the set of all probability distributions. A finite collection of real-valued continuous functions $\{f_i : i = 1, \dots, n\}$ and real constants $\{c_i : i = 1, \dots, n\}$ are given, and the corresponding *moment class* $\mathbb{P} \subset \mathcal{M}$ is defined to be the set of distributions satisfying the linear constraints,

$$\mathbb{P} := \{\pi \in \mathcal{M} : \langle \pi, f_i \rangle \leq c_i, \quad i = 1, \dots, n\}, \quad (1)$$

where the notation $\langle \pi, f \rangle$ is used to denote the mean of the function f according to the distribution π . It is assumed that the constraints are such that each $\pi \in \mathbb{P}$ is a probability distribution. We also consider equality constraints, using the same notation,

$$\mathbb{P} := \{\pi \in \mathcal{M} : \langle \pi, f_i \rangle = c_i, \quad i = 1, \dots, n\}, \quad (2)$$

However, it is convenient to transform the equality constraints to the form (1) to unify the presentation below.

The set of feasible moment vectors, and the set of feasible moment vectors in \mathbb{P} as defined in (1) are denoted,

$$\begin{aligned} \Delta &:= \{\langle \pi, f \rangle : \pi \in \mathcal{M}_1\} \subset \mathbb{R}^n \\ \Delta(f, c) &:= \Delta \cap \{x \in \mathbb{R}^n : x_i \leq c_i, i = 1, \dots, n\}. \end{aligned}$$

The following non-degeneracy condition is imposed throughout the paper:

- (A1)** The intersection $\Delta(f, c) \cap \text{int}(\Delta)$ is non-empty, where $\text{int}(\Delta)$ denotes the relative interior of Δ .

There are many applications in which the distribution π is not completely specified *a priori*. Our motivation for considering moment classes to model this uncertainty comes firstly from the simple observation that the most common approach to partial statistical modeling is through moments, typically mean and correlation. However, note that the functions $\{f_i\}$ do not have to be polynomials.

Probabilistic inference using moment information has a long and rich history [30, 4, 36, 18, 17, 21]. The primary motivation in these references comes from the fact that it is

possible to obtain worst-case bounds on the probability of a given set, over all probability distributions in a given moment class.

An example is digital communication over a wireless channel subject to severe time varying interference with a few dominant interferers (a situation that might arise in a CDMA system [38]). Here the additive noise term may not be well modeled as Gaussian or even statistically time-invariant. Training to learn the statistics may only result in partial knowledge of the marginal distribution π .

In the applications considered below the functional $I: \mathcal{M} \rightarrow [0, \infty]$ to be optimized can be chosen to be concave, so that the following optimization is a convex program,

$$\mathbf{max} \ I(\pi) \quad \mathbf{s. t.} \ \pi \in \mathbb{P}. \quad (3)$$

One can apply the Kuhn-Tucker alignment conditions to characterize an optimizer. Under general conditions this is expressed as follows: If π^* is an optimizer of (3), and $g_{\pi^*}: \mathbf{X} \rightarrow \mathbb{R}$ denotes the gradient, defined so that

$$I(\pi) \leq I(\pi^*) + \langle \pi - \pi^*, g_{\pi^*} \rangle, \quad \pi \in \mathcal{M},$$

then for some constants $\{\lambda_i\} \subset \mathbb{R}$,

$$\begin{aligned} g_{\pi^*}(x) &\leq f_\lambda(x), \quad x \in \mathbf{X}; \\ g_{\pi^*}(x) &= f_\lambda(x), \quad a.e. [\pi^*] \end{aligned} \quad (4)$$

where $f_\lambda = \sum \lambda_i f_i$.

Sometimes we are so fortunate to find that the functional I is linear, or can be closely approximated by a linear function. In such cases we write the linear program as

$$\mathbf{max} \ \langle \pi, g_0 \rangle \quad \mathbf{s. t.} \ \pi \in \mathbb{P} \quad (5)$$

where $g_0 \in C(\mathbf{X})$ is a continuous function on \mathbf{X} . An optimizer π^* for (5), if it exists, can be chosen so that it has at most $n + 2$ points of support. An optimizer of this form is called *extremal* [30, 27].

This paper provides a survey of applications involving special cases of these optimization problems. The common theme is that optimizers typically have finite support even in the non-linear program (3). This fact has significant value in the construction of algorithms for detection or coding in the applications considered.

II. CHANNEL CODING

One of the great success stories in communication theory is Shannon's celebrated 1948 paper [34] that demonstrates optimality of the Gaussian input distribution in the i.i.d. additive white Gaussian noise (AWGN) channel. It is now known that this is a very special case. There is an increasing list of channel models in which an optimizing distribution is *discrete*, with *finite* support. Examples include,

Gaussian channels It is shown in [35] that if the input is not only constrained by average power but also limited by a given peak power constraint, then the optimal input distribution has finite support. The conclusion of [35] is generalized to complex and vector Gaussian channels in [32, 24, 7].

Fading channels The capacity-achieving distribution for the Rayleigh channel is discrete in magnitude with a finite number of mass points, one of them located at the origin [1]. Similar conclusions hold for many other fading channel models [20, 11, 24]. Extensions to MIMO channels are contained in [23].

Worst-case channel modeling Discrete distributions 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 [33].

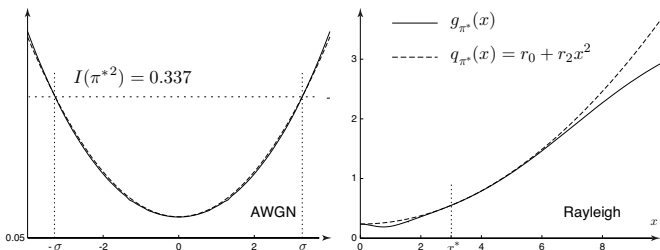


Figure 1: The illustration at left shows the alignment condition for the real AWGN channel, with $\pi = \frac{1}{2}(\delta_\sigma + \delta_{-\sigma})$. The illustration at right shows the alignment condition for the complex Rayleigh channel, with input distribution symmetric, with magnitude supported on $\{0, 5\}$.

Consider the channel capacity problem with real input and output alphabets, and transition density defined by

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

For a given average power constraint $\sigma_P^2 < \infty$, the channel capacity is expressed as the value of the nonlinear program (3) in which the set \mathbb{P} is defined in (2) with $\{f_i\} = \{1, x^2\}$, and $\{c_i\} = \{1, \sigma_P^2\}$. The functional $I: \mathcal{M} \rightarrow [0, \infty]$ to be optimized is the mutual information,

$$I(\pi) = \int \left(\int \ln \left(\frac{p(y|x)}{p(y|\pi)} \right) p(y|x) dy \right) \pi(dx), \quad \pi \in \mathcal{M}.$$

The alignment condition itself can rule out the existence of a continuous optimizer, which provides one explanation for the large number of examples in which π^* is discrete. This is best illustrated through example.

It can be shown that the gradient of I at π is equal to the *channel sensitivity function*, defined by the divergence

$$g_\pi(x) := D(p(\cdot|x) \| p(\cdot|\pi)) = \int \ln [p(y|x)/p(y|\pi)] p(y|x) dy.$$

The alignment condition (4) is expressed as follows: π^* is an optimizer if and only if there exists a quadratic function $q(x) = \lambda_0 + \lambda_2 x^2$ such that

$$\begin{aligned} g_{\pi^*}(x) &\leq q(x), \quad x \in \mathbf{X}; \\ g_{\pi^*}(x) &= q(x), \quad \text{a.e. } [\pi^*] \end{aligned}$$

Consider for example the real AWGN channel, and let π^{*2} denote the symmetric binary distribution supported on $\{-\sigma_P, \sigma_P\}$. Shown at left in Figure 1 is the sensitivity function $g_{\pi^{*2}}$ and a quadratic function $q_{\pi^{*2}}$ satisfying $g_{\pi^{*2}} \leq q_{\pi^{*2}}$ on $[-\sigma_P, \sigma_P]$ with $g_{\pi^{*2}}(\sigma_P) = q_{\pi^{*2}}(\sigma_P)$. It follows that the alignment condition holds for the convex program with peak power constraint given by σ_P . That is, the binary distribution π^{*2} is optimal distribution when the input alphabet is equal to the bounded set $\mathbf{X} = [-\sigma_P, \sigma_P]$.

The illustration at right in Figure 1 shows the alignment condition for the Rayleigh channel model in which a binary distribution is optimal without a peak power constraint. Further details may be found in [14].

When the average power constraint is significant, so that the signal to noise ratio (SNR) is small, then it is argued in [14] that the nonlinear program can be approximated by the linear program (5), where $g_0(x) := D(p(\cdot|x) \| p(\cdot|0))$. If an optimizer π^* of (5) exists, then without loss of generality it contains at most two points of support on \mathbf{X} (i.e. π^* is binary.) Similarly, Gallager in [10] showed that for small SNR, the reliability function for a finite-alphabet channel can be obtained by approximating a convex functional on \mathcal{M} by a linear functional. In this way the error-exponent optimization problem is transformed to a linear program of the form (5).

III. LARGE DEVIATIONS

Convex and linear optimization problems also arise in the following worst-case large-deviations problem. Consider first the classical LDP limit theorem for an i.i.d. sequence of random variables $\{X_j\}_{j=1}^\infty$ taking values in the compact set $\mathbf{X} \subset \mathbb{R}$, with marginal distribution $\pi \in \mathcal{M}_1$:

Theorem 1 (Sanov's Theorem for Empirical Measures) *The sequence of empirical distributions defined by*

$$L_N := \frac{1}{N} \sum_{j=0}^{N-1} \delta_{X_j}, \quad N \geq 1,$$

satisfies an LDP with respect to the weak-topology on \mathcal{M}_1 , with the good, convex rate-function*

$$J(\mu) := D(\mu \| \pi), \quad \mu \in \mathcal{M}_1.$$

Consequently, for any $E \in \mathcal{B}(\mathcal{M}_1)$,

$$\begin{aligned} - \inf_{\mu \in E^\circ} J(\mu) &\leq \liminf_{N \rightarrow \infty} N^{-1} \log L_N(E) \\ &\leq \limsup_{N \rightarrow \infty} N^{-1} \log L_N(E) \leq - \inf_{\mu \in \bar{E}} J(\mu), \end{aligned}$$

where E° and \bar{E} denote the interior and the closure of E in the weak-topology, respectively.* \square

The worst-case version of Sanov's Theorem is defined with respect to a given moment class: Suppose that π is not known, but is known to belong to the moment class \mathbb{P} . Then, the *worst-case rate-function* $L: \mathcal{M}_1 \rightarrow \mathbb{R}$ defined as,

$$L(\mu) := \min_{\pi \in \mathbb{P}} D(\mu \| \pi). \quad (6)$$

The optimization problem (6) is precisely of the form (3) with $I := -L$.

Theorem 2 is taken from [25, 27]. Part (ii) is a version of the alignment condition (4).

Theorem 2 (Worst-Case Sanov Bound) *The following hold under Assumption (A1):*

- (i) *The function L is convex; continuous in the weak*-topology; and uniformly bounded over \mathcal{M}_1 . Moreover, it has the representation,*

$$L(\mu) = \sup_{\lambda \in R_+(f)} \{ \langle \mu, \log \lambda^T f \rangle + 1 - \lambda^T c \}, \quad (7)$$

where

$$R_+(f) := \{ \lambda \in \mathbb{R}_+^{n+1} : \lambda^T f(x) \geq 0 \text{ for all } x \in X \}.$$

- (ii) *The infimum in (6) and the supremum in (7) are achieved by a pair $\pi^* \in \mathbb{P}$, $\lambda^* \in R_+(f)$, satisfying*

$$\frac{d\mu}{d\pi^*} = \lambda^{*T} f.$$

□

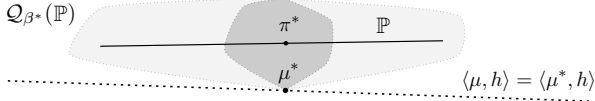


Figure 2: Geometric interpretation of extremal distributions. The dark region inside $\mathcal{Q}_{\beta^*}^+(\mathbb{P})$ is the divergence set $\mathcal{Q}_{\beta^*}^+(\pi^*)$.

Consider the special case in which $X = [0, 1]$, and the moment class is defined by the equality constraints (2) with $\{f_i\} = \{1, x, \dots, x^n\}$ the first n polynomials, and $c_i = \langle \nu, x^i \rangle = 100^{-1} \sum_{k=1}^{100} (k/100)^i$, $i \geq 1$. The vector c is consistent with the uniform distribution on X .

Shown in Figure 3 are results from numerical calculation of $L(\mu)$ for $n = 1, 3, 10$ and 20 with μ being the symmetric binary distribution supported on $\{0.01, 0.99\}$. Note that $\mu \in \mathbb{P}$ for $n = 1$, and hence $L(\mu) = 0$. Also shown in Figure 3 is the n th order polynomial $\lambda^T f$ in each case, and the roots of this polynomial contained in X . From Theorem 2 we know that $L(\mu) = D(\mu \parallel \pi^*)$ for some $\pi^* \in \mathbb{P}$ with $\frac{d\mu}{d\pi^*} = \lambda^T f$. It follows that π^* is supported on the union,

$$\text{supp}(\pi^*) \subset \{ \text{roots of } \lambda^T f \} \cup \{ \text{supp}(\mu) = \{0.1, 0.99\} \}$$

Consider now the LDP theorem in one dimension: For a given continuous function $h \in C(X)$, and a given marginal distribution $\pi \in \mathcal{M}_1$, Chernoff's bound for the exceedance probability is given by

$$\mathbb{P} \left\{ N^{-1} \sum_{j=1}^N h(X_j) \geq r \right\} \leq \exp(-N J_{\pi, h}(r)), \quad N \geq 1, \quad (8)$$

where $J_{\pi, h}(r)$ is the one-dimensional rate function,

$$J_{\pi, h}(r) = \inf \{ D(\mu \parallel \pi) : \mu \in \mathcal{M}_1 \text{ s.t. } \langle \mu, h \rangle \geq r \}. \quad (9)$$

If π is not known than we consider the worst-case, which requires that we minimize $J_{\pi, h}(r)$ over all $\pi \in \mathbb{P}$. Consider for $r \in \mathbb{R}$,

$$\mathcal{H} := \{ \mu \in \mathcal{M}_1 : \langle \mu, h \rangle = r \}, \quad (10)$$

$$\mathcal{H}^0 := \{ \mu \in \mathcal{M}_1 : \langle \mu, h \rangle < r \}, \quad (11)$$

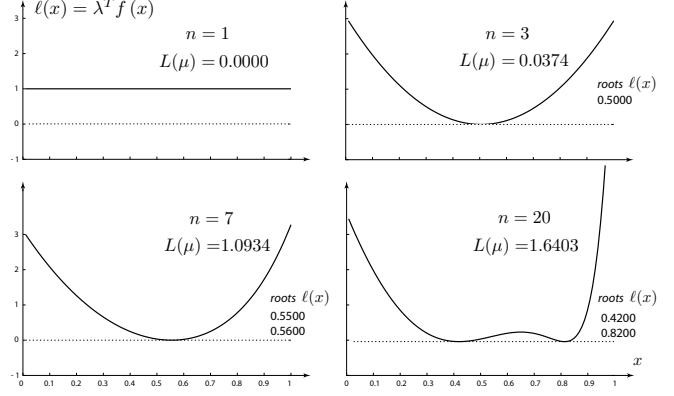


Figure 3: Computation of $L(\mu)$. Here μ is the symmetric binary distribution supported on $\{0.01, 0.99\}$, and n is the number of moment constraints used to define \mathbb{P} .

and $\mathcal{H}^1 := \{ \mu \in \mathcal{M}_1 : \langle \mu, h \rangle > r \}$. The set \mathcal{H} is an intersection of \mathcal{M}_1 and the hyperplane $\{ \mu \in \mathcal{S} : \langle \mu, h \rangle = r \}$, where \mathcal{S} denotes the set of signed measures on X . The set \mathcal{H} is closed in the weak* topology since h is continuous. Since it causes no ambiguity, we refer to \mathcal{H} itself as a hyperplane, and we refer to the sets $\{\mathcal{H}^0, \mathcal{H}^1\}$ as half-spaces. Based on Theorem 2 and the contraction principle we obtain a formula for the one-dimensional worst-case rate-function:

$$\begin{aligned} \underline{J}_h(r) &= \inf \{ D(\mu \parallel \pi) : \pi \in \mathbb{P}, \text{ and } \mu \in \mathcal{M}_1 \text{ s.t. } \langle \mu, h \rangle \geq r \} \\ &= \inf \{ L(\mu) : \mu \in \mathcal{M}_1 \text{ s.t. } \langle \mu, h \rangle \geq r \}, \quad r \in \mathbb{R}. \end{aligned} \quad (12)$$

Given a moment class \mathbb{P} , a function $h \in C(X)$, and $r \in \mathbb{R}$, a distribution $\pi^* \in \mathbb{P}$ is called $(h, r, +)$ -extremal if it solves the optimization (12) with $\underline{J}_h(r) > 0$. The '+' refers to the use of an upper tail in (8). A $(h, r, -)$ -extremal distribution is defined analogously.

A geometric interpretation of the extremal property is provided by convexity of the following divergence sets,

$$\mathcal{Q}_{\beta}^+(\pi) := \{ \mu \in \mathcal{M}_1 : D(\mu \parallel \pi) \leq \beta \}, \quad (13)$$

$$\mathcal{Q}_{\beta}^+(\mathbb{P}) := \bigcup_{\pi \in \mathbb{P}} \mathcal{Q}_{\beta}^+(\pi). \quad (14)$$

Divergence sets are convex subsets of \mathcal{M}_1 since $D(\cdot \parallel \cdot)$ is jointly convex [8, Theorem 2.7.2]. The minimization (12) may be expressed,

$$\underline{J}_h(r) = \inf_{\pi \in \mathbb{P}} \inf_{\mu \in \mathcal{H} \cup \mathcal{H}^1} D(\mu \parallel \pi), \quad r \in \mathbb{R}, \quad (15)$$

which is equivalently expressed in terms of divergence sets as,

$$\underline{J}_h(r) = \sup \{ \beta : \mathcal{Q}_{\beta}^+(\mathbb{P}) \cap \mathcal{H} = \emptyset \}, \quad r \in \mathbb{R}. \quad (16)$$

This geometry is illustrated in Figure 2.

The set \mathcal{H} forms a supporting hyperplane for $\mathcal{Q}_{\beta^*}^+(\mathbb{P})$, passing through distributions μ^* in the intersection $\mathcal{Q}_{\beta^*}^+(\mathbb{P}) \cap \mathcal{H}$. Theorem 2 asserts that there exists $\pi^* \in \mathbb{P}$ such that $D(\mu^* \parallel \pi^*) = \beta^*$. The pair of probability distributions $\{\mu^*, \pi^*\}$ solve (15), and π^* is a $(h, r, +)$ -extremal distribution.

The extremal property defined in this section is consistent with the definition of extremal distributions given previously: It is shown in [27] that $\pi^* \in \mathbb{P}$ is a $(r, h, +)$ -extremal distribution if and only if it is an optimizer of the linear program (5)

with $g_0 = e^{\theta^* h}$ for some $\theta^* > 0$. The value of the linear program in this case is an evaluation of the *worst-case moment generating function*, defined by

$$\begin{aligned} \max \langle \pi, \exp(\theta h) \rangle \\ \text{s. t. } \langle \pi, f_i \rangle \leq c_i, \quad i = 1, \dots, n, \quad \pi \in \mathcal{M}. \end{aligned} \quad (17)$$

An important special case of (17) is obtained on setting $\mathbf{X} = [0, 1]$, $h(x) \equiv x$, and \mathbb{P} is defined by the equality constraints (2) with $\{f_i\} = \{1, x, \dots, x^n\}$. The linear program (17) then reduces to a well-studied problem that originated in the nineteenth century work of A. A. Markov [21]. In this case, there is a single probability distribution $\pi^* \in \mathbb{P}$ that optimizes (17) simultaneously for every $\theta \in \mathbb{R}_+$. The optimizer π^* is known as a *Markov canonical distribution*.

The case $n = 1$ was considered by Hoeffding [12], and the case $n = 2$ was considered by Bennett [2] to obtain celebrated probability inequalities for sums of bounded random variables. In these two special cases the optimizing distribution π^* is binary. Since then, these ideas have been developed in various directions [2, 12, 22, 19, 39, 9, 31, 28, 30, 3, 36, 27, 25].

IV. ROBUST HYPOTHESIS TESTING

We now consider the binary hypothesis testing problem based on a finite number of observations from a sequence of observations $\mathbf{X} = \{X_t : t = 1, \dots\}$, taking values in \mathbf{X} . Throughout the remainder of the paper it is assumed that \mathbf{X} is finite.

Conditioned on either of 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 π^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.

Standard approaches to deciding whether the observations come from H_0 or H_1 include the Bayesian, Neyman-Pearson, and min-max criteria (see, e.g., [29]). It is well-known that when the distributions π^0 and π^1 are specified, the optimal test under any one of these three criteria can be expressed as a Likelihood Ratio Test (LRT) [29].

A reasonable way to capture partial knowledge of π^0 and π^1 is through sets of distributions referred to as *uncertainty classes*. A standard approach to designing decision rules in this setting is the min-max approach, where the goal is to minimize the worst-case performance over the uncertainty classes. The decision rules thus obtained are said to be robust to the uncertainties in the probability distributions. Min-max robust detection has been the subject of numerous papers since the seminal work of Huber and Strassen [15, 16]. The solution to the robust detection problem, if one exists, is a LRT between a pair of *least favorable distributions* (LFDs) within the classes. Huber and Strassen showed that LFDs exist for several uncertainty models that can be described generally in terms of alternating capacities of order 2 [16].

Motivated by the results surveyed in the previous section, we consider here uncertainty classes obtained by specifying bounds on a finite number of moments of the distributions under the respective hypotheses. Specifically, we define the two moment classes \mathbb{P}_0 and \mathbb{P}_1 as,

$$\mathbb{P}_j = \left\{ \pi \in \mathcal{M}_1 : \langle \pi, f_i \rangle \leq c_i^j, i = 1, \dots, n \right\}, \quad j = 0, 1, \quad (18)$$

where $\{f_i\}$ are real-valued continuous functions on \mathbf{X} , and $\{c_i^j\}$ are constants. It is assumed throughout that the sets

$\mathbb{P}_0, \mathbb{P}_1$ are disjoint, and that each satisfies the non-degeneracy assumption (A1).

Unfortunately, for uncertainly classes defined by moment constraints, the min-max robust versions of the standard hypothesis testing problems do not fall into the Huber and Strassen framework. To facilitate analysis, we turn to the asymptotic setting that is described in more detail below. Throughout this paper we restrict our attention to the asymptotic version of the Neyman-Pearson (N-P) criterion for evaluating a given detector. The results of this paper can be extended to asymptotic robust versions of the Bayesian and min-max hypothesis testing problems.

The asymptotic robust Neyman-Pearson (N-P) criterion is described as follows when the marginals π^0, π^1 are given [13]: Suppose that for each $N \geq 1$ a decision test ϕ_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,

$$\begin{aligned} I_\phi^{\pi^1} &:= - \liminf_{N \rightarrow \infty} \frac{1}{N} \log(\mathbb{P}_1(\phi_N(X_1, \dots, X_N) = 0)), \\ J_\phi^{\pi^0} &:= - \liminf_{N \rightarrow \infty} \frac{1}{N} \log(\mathbb{P}_0(\phi_N(X_1, \dots, X_N) = 1)), \end{aligned}$$

where $\pi^0, \pi^1 \in \mathbb{P}_0, \mathbb{P}_1$ are the actual marginal distributions of \mathbf{X} under H_0 and H_1 respectively, and $\{\mathbb{P}_i\}$ the corresponding distributions on sample space.

Consider first the non-robust setting in which the distributions π^0, π^1 are known exactly. The asymptotic N-P criterion for choosing an optimal test is to maximize the type-II exponent subject to a constraint on the type-I exponent. Thus, for a given constant $\eta \geq 0$ as the constraint, we have

$$\sup_{\phi} I_\phi^{\pi^1} \quad \text{subject to} \quad J_\phi^{\pi^0} \geq \eta \quad (19)$$

where the supremum is over all test sequences ϕ . The optimal value of the exponent $I_\phi^{\pi^1}$ in the asymptotic N-P problem is expressed in Theorem 3 below, which is a combination of results established in [13] and [5]. A sequence of tests that is asymptotically optimal is expressed using the log-likelihood ratio between π^0 and π^1 .

Given the two distributions $\pi^0, \pi^1 \in \mathcal{M}_1$ we define the likelihood ratio $\ell : \mathbf{X} \rightarrow \mathbb{R}$ by

$$\ell(x) = \frac{\pi_x^0}{\pi_x^1}, \quad x \in \mathbf{X}.$$

Theorem 3 *Suppose that the two distributions $\pi^0, \pi^1 \in \mathcal{M}_1$ are known exactly. Then the following statements hold,*

- (i) *The optimal value of $I_\phi^{\pi^1}$ in (19) is given by the minimal Kullback-Leibler divergence,*

$$\inf_{\mu \in \mathcal{Q}_\eta^+(\pi^0)} D(\mu \parallel \pi^1)$$

- (ii) *The minimization in (i) is uniquely attained by the distribution $\mu^* \in \mathcal{Q}_\eta^+(\pi^0)$ defined for some $s^* \geq 0$ and all $x \in \mathbf{X}$ by,*

$$\mu_x^* = (\pi_x^0)^{\frac{s^*}{1+s^*}} (\pi_x^1)^{\frac{1}{1+s^*}} \left(\sum_{x \in A} (\pi_x^0)^{\frac{s^*}{1+s^*}} (\pi_x^1)^{\frac{1}{1+s^*}} \right)^{-1}.$$

Moreover, we must have either $s^* = 0$, or $D(\mu^* \parallel \pi^0) = \eta$ (or both).

(iii) Suppose that $D(\mu^* \parallel \pi^0) = \eta$, and define $\beta^* = D(\mu^* \parallel \pi^1)$. Then, the log-likelihood ratio defines a LRT test ϕ^* that is asymptotically optimal: In this test the decision region for H_1 is given by,

$$A_1(N) := \left\{ x \in \mathcal{X}^N : \frac{1}{N} \sum_{t=1}^N \log \ell(x_t) \leq \beta^* - \eta \right\}. \quad (20)$$

□

Part (i) of Theorem 3 was proved by Hoeffding [13]. Parts (ii) and (iii) were established in [5] for the case in which the support of π^0 and π^1 is all of \mathcal{X} . A sketch of the proof is provided in the Appendix.

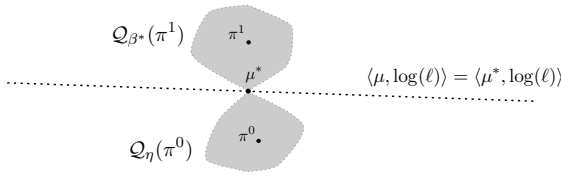


Figure 4: The Neyman-Pearson hypothesis testing problem. The likelihood ratio test is interpreted as a separating set between the convex sets $\mathcal{Q}_{\eta}^+(\pi^0)$ and $\mathcal{Q}_{\beta^*}^+(\pi^1)$.

Theorem 3 may be interpreted geometrically as follows. On setting

$$\beta^* = \sup\{\beta \geq 0 : \mathcal{Q}_{\eta}^+(\pi^0) \cap \mathcal{Q}_{\beta}^+(\pi^1) = \emptyset\} = \inf_{\mu \in \mathcal{Q}_{\eta}^+(\pi^0)} D(\mu \parallel \pi^1),$$

we have $\mu^* \in \mathcal{Q}_{\eta}^+(\pi^0) \cap \mathcal{Q}_{\beta^*}^+(\pi^1)$. 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 (20):

$$\mathcal{H} = \{\mu \in \mathcal{M}_1 : \langle \mu, \log \ell \rangle = \langle \mu^*, \log \ell \rangle\}$$

This geometry is illustrated in Figure 4.

We now present an extension of Theorem 3 to the robust framework in which π^0 and π^1 are known to belong to the uncertainty classes \mathbb{P}_0 and \mathbb{P}_1 respectively. We impose a uniform constraint on the type-I exponent for all $\pi^0 \in \mathbb{P}_0$, and subject to this we seek a test sequence that maximizes the worst type-II exponent across $\pi^1 \in \mathbb{P}_1$. Thus, in the robust version, the asymptotic N-P criterion (19) is replaced by the following constrained optimization:

$$\sup_{\phi} \inf_{\pi^1 \in \mathbb{P}_1} I_{\phi}^{\pi^1} \quad \text{subject to} \quad \inf_{\pi^0 \in \mathbb{P}_0} J_{\phi}^{\pi^0} \geq \eta. \quad (21)$$

Theorem 4 establishes the existence of an optimal test sequence ϕ^* in which a log-linear combination of the constraint functions $\{f_i : 0 \leq i \leq n\}$ is compared to a threshold.

The proof of Theorem 4 is given in [26]. Assumption (22) is imposed to ensure that the solution to the robust hypothesis testing problem is non-trivial.

Theorem 4 Consider the asymptotic robust N-P hypothesis testing problem (21) under Assumption (A1), and the additional assumption,

$$D(\pi^1 \parallel \pi^0) > \eta \text{ for each } \pi^0 \in \mathbb{P}_0, \text{ and } \pi^1 \in \mathbb{P}_1. \quad (22)$$

That is, $\mathcal{Q}_{\eta}^+(\mathbb{P}_0) \cap \mathbb{P}_1 = \emptyset$. Then,

(i) The optimal value of the exponent in (21) is given by the minimal Kullback-Leibler divergence,

$$\begin{aligned} \beta^* &:= \sup\{\beta \geq 0 : \mathcal{Q}_{\eta}^+(\mathbb{P}_0) \cap \mathcal{Q}_{\beta}^+(\mathbb{P}_1) = \emptyset\} \\ &= \inf_{\pi^1 \in \mathbb{P}_1} \inf_{\mu \in \mathcal{Q}_{\eta}^+(\mathbb{P}_0)} D(\mu \parallel \pi^1). \end{aligned} \quad (23)$$

(ii) There exist $\pi^{0*} \in \mathbb{P}_0, \pi^{1*} \in \mathbb{P}_1$, and a probability distribution μ^* that solve (23). The distributions π^{0*} and π^{1*} are mutually absolutely continuous and, for some $s^* \geq 0$, on their support

$$\frac{\pi_x^{0*}}{\pi_x^{1*}} = C_0 \ell_0^{(1+\frac{1}{s^*})}(x) = C_1^{-1} \ell_1^{-(1+s^*)}(x),$$

where the functions ℓ_0 and ℓ_1 are given by $\ell_0 = \lambda^T f$ and $\ell_1 = \gamma^T f$. Moreover, μ^* may be expressed as $\mu_x^* = \ell_1(x)\pi_x^{0*} = \ell_0(x)\pi_x^{1*}$.

(iii) An optimal sequence of tests is defined through the decision regions

$$A^1(N) = \left\{ x \in \mathcal{X}^N : \frac{1}{N} \sum_{t=1}^N \log \ell^*(x_t) \leq \beta^* - \eta \right\},$$

where $\ell^*(x) := \ell_0(x)/\ell_1(x)$ for $x \in \mathcal{X}$.

□

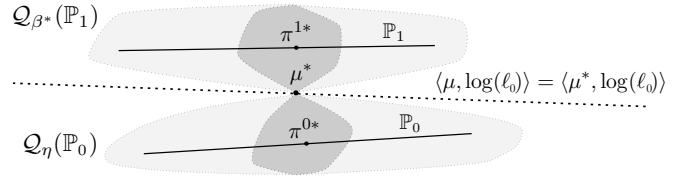


Figure 5: The two-moment worst-case hypothesis testing problem. The uncertainty classes \mathbb{P}_i , $i = 0, 1$ are determined by a finite number of linear constraints, and the thickened regions $\mathcal{Q}_{\eta}^+(\mathbb{P}_0)$, $\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$ are each convex. The linear threshold test is interpreted as a separating hyperplane between these two convex sets.

Theorem 4 has a geometric interpretation that is entirely analogous to that given for Theorem 3. With $\beta^* \geq 0$ defined in (23), we find that

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

and that $\mu^* \in \mathcal{Q}_{\eta}^+(\mathbb{P}_0) \cap \mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$. The probability distributions $\{\pi^{0*}, \pi^{1*}\}$ belong to the respective moment classes, and satisfy,

$$D(\mu^* \parallel \pi^{0*}) = \eta \text{ and } D(\mu^* \parallel \pi^{1*}) = \beta^*.$$

From Theorem 3, β^* is clearly an upper bound on the optimal exponent in (21). Theorem 4 shows that this bound is attained. Either of the functions $\log \ell_0$ or $\log \ell_1$ defined in the theorem 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 5.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Award Nos. ECS 02-17836, ITR 00-85929 and the PECASE award CCF 00-49089. Meyn and Huang express their thanks to Prof. M. Medard at MIT for many valuable conversations on channel coding.

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] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.
- [3] D. Bertsimas and J. Sethuraman. Moment problems and semidefinite optimization. In *Handbook of semidefinite programming*, volume 27 of *Internat. Ser. Oper. Res. Management Sci.*, pages 469–509. Kluwer Acad. Publ., Boston, MA, 2000.
- [4] D. Bertsimas and J. Sethuraman. Moment problems and semidefinite optimization. In *Handbook of semidefinite programming*, volume 27 of *Internat. Ser. Oper. Res. Management Sci.*, pages 469–509. Kluwer Acad. Publ., Boston, MA, 2000.
- [5] R. E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Information Theory*, IT-20:405–417, 1974.
- [6] J.-F. Chamberland and V. V. Veeravalli. Decentralized detection in sensor networks. *IEEE Transactions on Signal Processing*, 51:407–416, 2003.
- [7] T.H. Chan, S. Hranilovic, and F.R. Kschischang. Capacity-achieving probability measure for vector Gaussian channels with bounded inputs - part I: Signal-independent noise. Submitted for publication, *IEEE Trans. Inform. Theory*.
- [8] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons Inc., New York, 1991. A Wiley-Interscience Publication.
- [9] P. Diaconis. Application of the method of moments in probability and statistics. In *Moments in mathematics (San Antonio, Tex., 1987)*, volume 37 of *Proc. Sympos. Appl. Math.*, pages 125–142. Amer. Math. Soc., Providence, RI, 1987.
- [10] 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, Mass, 2002.
- [11] M.C. Gursoy, H.V. Poor, and S. Verdú. The noncoherent Rician fading channel - part I: Structure of capacity achieving input. To appear, *IEEE Trans. Wireless Communication*.
- [12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.
- [13] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, 36:369–408, 1965.
- [14] J. Huang and S. Meyn. Characterization and computation of optimal distributions for channel coding. To appear, *IEEE Trans. Info. Theory*.
- [15] P. J. Huber. A robust version of the probability ratio test. *Ann. Math. Statist.*, 36:1753–1758, 1965.
- [16] P. J. Huber and V. Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.*, 1:251–263, 1973.
- [17] Mary A. Johnson and Michael R. Taaffe. Matching moments to phase distributions: nonlinear programming approaches. *Comm. Statist. Stochastic Models*, 6(2):259–281, 1990.
- [18] Mary A. Johnson and Michael R. Taaffe. An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queueing Systems Theory Appl.*, 8(2):129–147, 1991.
- [19] S. Karlin and W. J. Studden. *Tchebycheff systems: With applications in analysis and statistics*. Pure and Applied Mathematics, Vol. XV. Interscience Publishers John Wiley & Sons, New York-London-Sydney, 1966.
- [20] M. Katz and S. Shamai. On the capacity-achieving distribution of the discrete-time non-coherent additive white Gaussian noise channel. In *IEEE International Symposium on Information Theory*, page 165, 2002.
- [21] M. G. Kreĭn. The ideas of P. L. Čebyšev and A. A. Markov in the theory of limiting values of integrals and their future developments. *Translations of the American Mathematical Society*, 12:1–121, 1959.
- [22] A. W. Marshall and I. Olkin. *Inequalities: theory of majorization and its applications*, volume 143 of *Mathematics in Science and Engineering*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1979.
- [23] 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.
- [24] R. Palanki. On the capacity-achieving distributions of some fading channels. In *Proceedings of 40th Allerton Conference on Communication, Control, and Computing*, 2002.
- [25] 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.
- [26] C. Pandit and S. P. Meyn and V. V. Veeravalli. Asymptotic Robust Neyman-Pearson Hypothesis Testing Based on Moment Classes. In preparation, 2004.
- [27] C. Pandit and S. P. Meyn. Extremal distributions and worst-case large-deviation bounds. Submitted for publication, 2004.
- [28] J. E. Pečarić, F. Proschan, and Y. L. Tong. *Convex functions, partial orderings, and statistical applications*, volume 187 of *Mathematics in Science and Engineering*. Academic Press Inc., Boston, MA, 1992.
- [29] H. V. Poor. *An introduction to signal detection and estimation*. Springer Texts in Electrical Engineering. Springer-Verlag, New York, second edition, 1994. A Dowden & Culver Book.
- [30] F. Popescu and D. Bertsimas. Optimal inequalities in probability theory: A convex optimization approach. INSEAD working paper TM62, <http://faculty.insead.edu/popescu/ioana/myresearch.htm>, 2003.
- [31] M. Shaked and Y. L. Tong, editors. *Stochastic inequalities*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 22. Institute of Mathematical Statistics, Hayward, CA, 1992.
- [32] S. Shamai and I. Bar-David. The capacity of average and peak-power-limited quadrature Gaussian channels. *IEEE Transactions on Information Theory*, 41(4):1060–1071, 1995.
- [33] S. Shamai and S. Verdú. Worst-case power-constrained noise for binary-input channels. *IEEE Transactions on Information Theory*, 38(5):1494–1511, 1992.
- [34] C.E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27:379–423, 1948.
- [35] J.G. Smith. The information capacity of amplitude and variance-constrained scalar Gaussian channels. *Inform. Contr.*, 18:203–219, 1971.
- [36] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyshev bounds via semidefinite programming. Submitted to SIAM Review, Problems and Techniques Section, January 2004.
- [37] V. V. Veeravalli, T. Basar, and H. V. Poor. Minimax robust decentralized detection. *IEEE Transactions on Information Theory*, 40(1):35–40, 1994.
- [38] A. J. Viterbi. *CDMA: Principles of Spread Spectrum Communications*. Addison-Wesley, New York, 1995.
- [39] W. Whitt. Bivariate distributions with given marginals. *Ann. Statist.*, 4(6):1280–1289, 1976.