Worst-Case Large-Deviations Asymptotics with Application to Queueing and Information Theory*

Charuhas Pandit† Sean Meyn‡

November 16, 2007

Abstract

An i.i.d. process $\boldsymbol{X}$ is considered on a compact metric space X. Its marginal distribution π is unknown, but is assumed to lie in a moment class of the form,

$$\mathbb{P} = \{\pi : \langle \pi, f_i \rangle = c_i, \quad i = 1, \ldots, n\},$$

where $\{f_i\}$ are real-valued, continuous functions on X, and $\{c_i\}$ are constants. The following conclusions are obtained:

(i) For any probability distribution μ on X, Sanov's rate-function for the empirical distributions of $\boldsymbol{X}$ is equal to the Kullback-Leibler divergence $D(\mu \parallel \pi)$. The worst-case rate-function is identified as

$$L(\mu) := \inf_{\pi \in \mathbb{P}} D(\mu \parallel \pi) = \sup_{\lambda \in R(f,c)} \left\langle \mu, \log(\lambda^T f) \right\rangle,$$

where $f = (1, f_1, \ldots, f_n)^T$, and $R(f, c) \subset \mathbb{R}^{n+1}$ is a compact, convex set.

(ii) A stochastic approximation algorithm for computing L is introduced based based on samples of the process $\boldsymbol{X}$.

(iii) A solution to the worst-case one-dimensional large-deviations problem is obtained through properties of *extremal distributions*, generalizing Markov's canonical distributions.

(iv) Applications to robust hypotheses testing and to the theory of buffer overflows in queues are also developed.

*This paper is based upon work supported by the National Science Foundation under Award Nos. ECS 02 17836 and ITR 00-85929. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation

†C. P. is with Morgan Stanley and Co., 1585 Broadway, New York, NY, 10019 (charuhas.pandit@morganstanley.com)

‡S. M. is with the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering at the University of Illinois, 1308 W. Main Street, Urbana, IL 61801, URL `http://decision.csl.uiuc.edu:80/~meyn` (meyn@uiuc.edu).

1 Introduction & Background

Consider an i.i.d. sequence $\boldsymbol{X}$ on a compact metric space X. It is assumed that its marginal distribution π is not known exactly, but belongs to the *moment class* $\mathbb{P}$ defined as follows: A finite set of real-valued continuous functions $\{f_i : i = 1, \dots, n\}$ and real constants $\{c_i : i = 1, \dots, n\}$ are given, and

$$\mathbb{P} := \{\pi \in \mathcal{M}_1 : \langle \pi, f_i \rangle = c_i, \quad i = 1, \dots, n\}, \tag{1}$$

where $\mathcal{M}_1$ is the space of probability distributions on X, and the notation $\langle \pi, f_i \rangle$ is used to denote the mean of the function f_i according to the distribution π.

The motivation for consideration of moment classes comes primarily from the simple observation that the most common approach to partial statistical modeling is through moments, typically mean and correlation. Moment classes have been considered in applications to finance [41]; admission control [10, 5, 32]; queueing theory [20, 19, 12]; and other applications.

For a moment class of this form, and a given function $g \in C(\mathsf{X})$, the map $\pi \to \langle \pi, g \rangle$ defines a continuous linear functional on $\mathcal{M}_1$. Consequently, the following maximization may be viewed as a linear program,

$$\max_{\pi \in \mathbb{P}} \langle \pi, g \rangle \tag{2}$$

The value of this (infinite-dimensional) linear program provides a bound on the mean of g that is uniform over $\pi \in \mathbb{P}$. A. A. Markov, a student of Chebyshev, considered a special case of the linear programs (2) in which the functions $\{f_i\}$ are polynomials. A comprehensive survey by M. G. Kreĭn in 1959 describes many of Markov's original results [26]. Since then, these ideas have been developed in various directions [1, 13, 44, 29, 21, 8, 38, 34, 36, 3, 42].

The present paper concerns various large-deviations bounds that are uniform across a moment class. One set of results concerns relaxations of *Chernoff's bound*: For a given function $h \in C(\mathsf{X})$, and any $r \geq \langle \pi, h \rangle$,

$$\mathsf{P}\{S_N \geq r\} \leq \exp(-N I_{\pi,h}(r)), \qquad N \geq 1, \tag{3}$$

where $\{S_N = N^{-1} \sum_{j=1}^N h(X_j) : N \geq 1\}$, and $I_{\pi,h}$ is the usual one-dimensional large deviations rate-function under the distribution π. Denoting the log moment-generating function as,

$$M_{\pi,h}(\theta) := \log \langle \pi, \exp(\theta h) \rangle, \qquad \theta \in \mathbb{R}, \tag{4}$$

the rate-function is equal to the convex dual,

$$I_{\pi,h}(r) = \sup_{\theta \in \mathbb{R}} \{\theta r - M_{\pi,h}(\theta)\}, \qquad r \in \mathbb{R}. \tag{5}$$

Theorem 1.5 and related results in Section 3 contain expressions for the minimum of the rate function $I_{\pi,h}(r)$ over the moment class $\mathbb{P}$. To put these results in context we present some known results in the special case of polynomial constraint functions.

1.1 Markov's canonical distributions

Suppose that $\mathsf{X} = [0, 1]$, $h(x) \equiv x$, and that the constraint functions $\{f_i\}$ are of the form,

$$f_i(x) = x^i, \qquad x \in \mathsf{X} = [0, 1], \ i = 1, \dots, n. \tag{6}$$

A direct approach is to introduce the *worst-case moment-generating function*, which for each $\theta \in \mathbb{R}$ is a special case of (2), defined as

$$\overline{m}_h(\theta) := \max_{\pi \in \mathbb{P}} \langle \pi, \exp(\theta h) \rangle, \qquad \theta \in \mathbb{R}. \tag{7}$$

The solution of this linear program gives a uniform lower bound on the rate-function (5). Under mild conditions on the vector c used in (1), it is shown in [26] that there is a single probability distribution $\pi^* \in \mathbb{P}$ that optimizes (7) simultaneously *for every* $\theta \in \mathbb{R}_+$. The probability distribution π^* is known as a *Markov canonical distribution*.

Theorem 1.1. (Markov's Canonical Distributions) *Suppose that h is the identity function on $[0,1]$; the functions $\{f_i\}$ are given in (6) for some $n \geq 1$; and that the vector $(c_1, \ldots, c_n)^T$ lies in the interior of the set of feasible moment vectors,*

$$\Delta := \{x \in \mathbb{R}^n : x_i = \langle \pi, f_i \rangle, i = 1, \ldots, n, \text{ for some } \pi \in \mathcal{M}_1\}. \tag{8}$$

Then,

(i) *There exists a probability distribution $\pi^* \in \mathbb{P}$, depending only on the moment constraints $\{c_i\}$, that optimizes the linear program (7) for each $\theta \geq 0$.*

(ii) *The probability distribution π^* is a discrete distribution with exactly $\lceil n/2 \rceil + 1$ points of support. Moreover, if n is even, then the end-point 1 lies in the support of π^*. If n is odd, then the end-points $\{0, 1\}$ each lie in the support of π^*.*

□

It can be shown that finding the distribution π^* is equivalent to solving an n^{th} degree polynomial. Consequently, analytical formulae for π^* are available for $n \leq 4$. Consider the following two special cases:

(i) *A single mean-constraint.* When $n = 1$, the canonical distribution is supported on 0 and 1, with $\pi^*(\{0\}) = 1 - \pi^*(\{1\}) = c_1$.

(ii) *First and second moment constraints.* The canonical distribution π^* is again binary when $n = 2$, and can be expressed

$$\pi^* = p_0 \delta_{x_0} + (1 - p_0)\delta_1, \tag{9}$$

where $x_0 = \frac{c_1 - c_2}{1 - c_1}$, and $p_0 = \frac{(1-c_1)^2}{1 + c_2 - 2c_1}$.

The case $n = 1$ was considered by Hoeffding [13], and the case $n = 2$ was considered by Bennett [1] to obtain celebrated probability inequalities for sums of bounded random variables. The following Generalized Bennett's Theorem follows directly from Theorem 1.1 and Chernoff's bound (3).

Theorem 1.2. (Generalized Bennett's Theorem) *Suppose that the assumptions of Theorem 1.1 hold. Consider the worst-case, one-dimensional rate-function defined by,*

$$\underline{I}(r) := \inf\{I_\pi(r) : \pi \in \mathbb{P}\}, \qquad r \in \mathbb{R}, \tag{10}$$

where the rate-function I_π is defined in (5) with $h(x) \equiv x$. Then, the Markov canonical distribution π^* achieves the point-wise minimum:

$$I_{\pi^*}(r) = \underline{I}(r), \qquad r \geq c_1. \tag{11}$$

Consequently, the universal Chernoff bound holds,

$$\mathsf{P}\{S_N \geq r\} \leq \exp(-NI_{\pi^*}(r)), \qquad \pi \in \mathbb{P},\ N \geq 1,\ r \geq c_1.$$

□

1.2 Main results

How can Theorems 1.1 and 1.2 be extended to allow general constraint functions, or a general compact state space? Theorem 1.4 provides generalizations in both directions, and also gives a transparent bound on the empirical distribution large-deviations asymptotics.

We first recall some well know definitions and results [7]. For two distributions $\mu, \pi \in \mathcal{M}_1$, the relative entropy, or Kullback-Leibler divergence is defined as,

$$D(\mu \parallel \pi) = \begin{cases} \langle \pi, \frac{d\mu}{d\pi} \log \frac{d\mu}{d\pi} \rangle & \text{if } \mu \prec \pi, \\ \infty & \text{otherwise} \end{cases}$$

The domain of definition of D is usually restricted to the space of probability distributions $\mathcal{M}_1$, but for the convex analytic methods to be applied in this paper, we extend the definition of D in the obvious way to include the space $\mathcal{M}$ of all finite positive measures on X.

We let $\mathcal{S}$ denote the set of signed measures on X with finite mass, so that $|\mu| \in \mathcal{M}$ for $\mu \in \mathcal{S}$. We assume that $\mathcal{S}$ is endowed with the weak*-topology, defined to be the smallest topology on $\mathcal{S}$ that contains the system of neighborhoods

$$\left\{\mu \in \mathcal{S} : |\mu(g) - s| < \epsilon\right\}, \quad \text{for real-valued } g \in C(\mathsf{X}),\ s \in \mathbb{R},\ \epsilon > 0\Big\}. \tag{12}$$

The associated Borel σ-field induced by the weak*-topology on $\mathcal{M}_1$ is denoted $\mathcal{F}$.

The sequence of *empirical distributions* is defined by,

$$L_N := \frac{1}{N} \sum_{j=0}^{N-1} \delta_{X_j}, \qquad N \geq 1. \tag{13}$$

We then have the well-known limit theorem [37, 7].

Theorem 1.3. (Sanov's Theorem for Empirical Measures) *Suppose that $\boldsymbol{X}$ is i.i.d. with marginal distribution π on the compact state space X. The sequence of empirical measures $\{L_N\}$ satisfies an LDP in the space $(\mathcal{M}_1, \mathcal{F})$ equipped with the weak*-topology, with the good, convex rate-function*

$$I(\mu) := D(\mu \parallel \pi), \qquad \mu \in \mathcal{M}_1. \tag{14}$$

Consequently, for any $E \in \mathcal{F}$,

$$\begin{aligned} -\inf_{\mu \in E^o} I(\mu) &\leq \liminf_{N \to \infty} N^{-1} \log L_N(E) \\ &\leq \limsup_{N \to \infty} N^{-1} \log L_N(E) \ \leq\ -\inf_{\mu \in \overline{E}} I(\mu)\,, \end{aligned}$$

where E^o and $\overline{E}$ denote the interior and the closure of E in the weak-topology, respectively.* □

On considering the special case $E = \{\mu \in \mathcal{M}_1 : \langle \mu, h \rangle \geq r\}$ for $r \in \mathbb{R}$, Theorem 1.3 implies the following representation of the one-dimensional rate-function,

$$I_{\pi,h}(r) = \inf\{D(\mu \parallel \pi) : \mu \in \mathcal{M}_1 \text{ s.t. } \langle \mu, h \rangle \geq r\} \tag{15}$$

Equation (15) is known as the contraction principle.

In view of Theorem 1.3 and the representation (15), we are led to seek *lower bounds* on the rate-function I defined in (14).

We are now in a position to state the main result of this paper. Theorem 1.4 provides an expression for the *worst-case rate-function* $L\colon \mathcal{M}_1 \to \mathbb{R}$ defined as,

$$L(\mu) := \inf_{\pi \in \mathbb{P}} D(\mu \parallel \pi). \tag{16}$$

For an arbitrary probability distribution $\pi \in \mathcal{M}_1$ and for $\beta \in \mathbb{R}_+$, the *divergence sets* $\mathcal{Q}_\beta(\pi), \mathcal{Q}_\beta^+(\pi)$ are defined as

$$\begin{aligned} \mathcal{Q}_\beta(\pi) &:= \{\mu \in \mathcal{M}_1 : D(\mu \parallel \pi) < \beta\}, \\ \mathcal{Q}_\beta^+(\pi) &:= \{\mu \in \mathcal{M}_1 : D(\mu \parallel \pi) \leq \beta\}. \end{aligned} \tag{17}$$

Divergence sets are convex subsets of $\mathcal{M}_1$ since $D(\cdot \parallel \pi)$ is a convex function. The above definition is extended to include divergence sets of the moment class $\mathbb{P}$:

$$\mathcal{Q}_\beta(\mathbb{P}) = \bigcup_{\pi \in \mathbb{P}} \mathcal{Q}_\beta(\pi) \qquad \text{and} \qquad \mathcal{Q}_\beta^+(\mathbb{P}) = \bigcup_{\pi \in \mathbb{P}} \mathcal{Q}_\beta^+(\pi). \tag{18}$$

We have $L(\mu) < \beta$ if and only if $\mu \in \mathcal{Q}_\beta(\mathbb{P})$.

The following assumptions on these constraint functions and constants are imposed throughout the paper:

(A1) The functions $1, f_1, \ldots, f_n$, are continuous on X, and the vector $(c_1, \ldots, c_n)^T$ lies in the interior of the set of feasible moment vectors, defined as

$$\Delta := \{x \in \mathbb{R}^n : x_i = \langle \pi, f_i \rangle, i = 1, \ldots, n, \text{ for some } \pi \in \mathcal{M}_1\}. \tag{19}$$

A version of Theorem 1.4 appears as Proposition 2.2.1 in the dissertation [31]. A proof is included in the Appendix.

Let $\{f_1, \ldots, f_n\}$ be the continuous functions and $\{c_i, \ldots, c_n\}$ the constants used in the definition (1). We let $f\colon \mathsf{X} \to \mathbb{R}^{n+1}$ denote the vector of functions $(1, f_1, \ldots, f_n)^T$, and write $c := (1, c_1, \ldots, c_n)^T \in \mathbb{R}^{n+1}$.

Theorem 1.4. (Worst-Case Sanov Bound) *The following hold under Assumption (A1):*

(i) *The function L may be expressed,*

$$L(\mu) = \sup_{\lambda \in R(f)} \{\langle \mu, \log \lambda^T f \rangle + 1 - \lambda^T c\}, \tag{20}$$

where

$$R(f) := \{\lambda \in \mathbb{R}^{n+1} \colon \lambda^T f(x) \geq 0 \text{ for all } x \in \mathsf{X}\}. \tag{21}$$

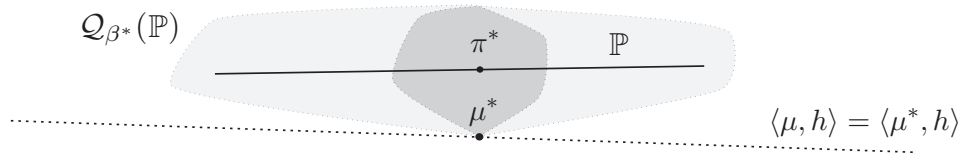

Figure 1: Geometric interpretation of extremal distributions. The dark region inside $\mathcal{Q}^+_{\beta^*}(\mathbb{P})$ is the divergence set $\mathcal{Q}^+_{\beta^*}(\pi^*)$.

(ii) *The infimum in (16) and the supremum in (20) are achieved by a pair* $\pi^* \in \mathbb{P}$, $\lambda^* \in R(f)$, *satisfying*

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

Consequently, $\lambda^{*T} c = 1$.

(iii) *The function* L *is convex; it is continuous in the weak*-topology; and it is uniformly bounded:*

$$\sup_{\mu \in \mathcal{M}_1} L(\mu) < \infty.$$

(iv) *For* $\beta > 0$, *the sets* $\mathcal{Q}_\beta(\mathbb{P}), \mathcal{Q}^+_\beta(\mathbb{P})$ *defined in (18) are convex. These sets also enjoy the following properties in the weak*-topology: The set* $\mathcal{Q}^+_\beta(\mathbb{P})$ *is compact, the set* $\mathcal{Q}_\beta(\mathbb{P})$ *is open, and the closure of* $\mathcal{Q}_\beta(\mathbb{P})$ *is equal to* $\mathcal{Q}^+_\beta(\mathbb{P})$. □

Let $C(\mathsf{X})$ denote the set of continuous functions on X, and define for $h \in C(\mathsf{X})$, $r \in \mathbb{R}$,

$$\mathcal{H} := \{\mu \in \mathcal{M}_1 : \langle \mu, h \rangle = r\}, \tag{22}$$

$$\mathcal{H}^0 := \{\mu \in \mathcal{M}_1 : \langle \mu, h \rangle < r\}, \qquad \text{and} \qquad \mathcal{H}^1 := \{\mu \in \mathcal{M}_1 : \langle \mu, h \rangle > r\}. \tag{23}$$

The set $\mathcal{H}$ is an intersection of $\mathcal{M}_1$ and the *hyperplane* $\{\mu \in \mathcal{S} : \langle \mu, h \rangle = r\}$. The set $\mathcal{H}$ is *closed* in the weak* topology since $h \in C(\mathsf{X})$. 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.

The function $\underline{I}_h$ is convex and non-negative on $\mathbb{R}$, it is identically zero on the interval $[\underline{r}_h, \overline{r}_h]$, and identically infinite on $[\underline{h}, \overline{h}]^c$, where

$$\begin{aligned} \overline{r}_h &= \sup\{r : \mathcal{H}(r) \cap \mathbb{P} \neq \emptyset\}, & \overline{h} &= \max\{h(x) : x \in \mathsf{X}\}; \\ \underline{r}_h &= \inf\{r : \mathcal{H}(r) \cap \mathbb{P} \neq \emptyset\}, & \underline{h} &= \min\{h(x) : x \in \mathsf{X}\}. \end{aligned} \tag{24}$$

An interpretation of these constants is illustrated in Figure 2, and in Proposition 3.4 below.

Based on Theorem 1.4 and the contraction principle (15), we obtain a formula for the worst-case rate-function in one-dimension on the closed interval $[\overline{r}_h, \overline{h}]$:

$$\begin{aligned} \underline{I}_h(r) &:= \inf_{\pi \in \mathbb{P}} I_{\pi,h}(r) \\ &= \inf\{D(\mu \,\|\, \pi) : \pi \in \mathbb{P}, \quad \text{and } \mu \in \mathcal{M}_1 \text{ s.t. } \langle \mu, h \rangle \geq r\} \\ &= \inf\{L(\mu) : \mu \in \mathcal{M}_1 \text{ s.t. } \langle \mu, h \rangle \geq r\}, \qquad r \in [\overline{r}_h, \overline{h}]. \end{aligned} \tag{25}$$

This gives rise to the notion of extremal distributions:

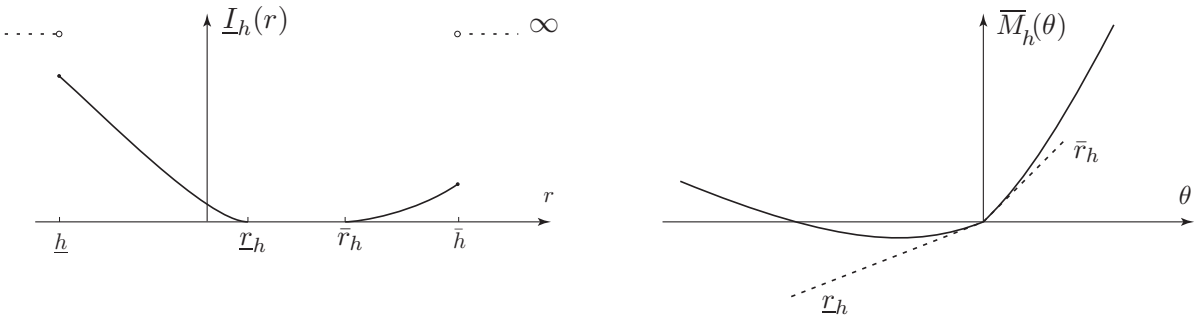

Figure 2: Plot of a typical worst-case one-dimensional rate-function, and worst-case log moment-generating function.

> Given a moment class $\mathbb{P}$, a function $h \in C(\mathsf{X})$, and $r \in (\overline{r}_h, \overline{h})$, a distribution $\pi^* \in \mathbb{P}$ is called $(h, r, +)$-*extremal* if it solves the optimization (25).

The '+' refers to the use of an *upper tail* in (3). The constraint $r \in (\overline{r}_h, \overline{h})$ ensures that $\underline{I}_h(r) > 0$. A $(h, r, -)$-*extremal* distribution is defined analogously for $r \in (\underline{h}, \underline{r}_h)$.

When the precise values of h and r are unimportant, we simply refer to π^* as an extremal distribution.

The paper [36] uses the exact same terminology for distributions that solve a particular infinite dimensional linear program. Although the setting is very different, Theorem 1.5 shows that the definition used here is consistent with the definition of extremal distributions introduced in [36]. A proof is provided in the Appendix..

Theorem 1.5. (Saddle-Point Property) *For any $r \in (\overline{r}_h, \overline{h})$, there exists $\pi^* \in \mathbb{P}$ and $\theta^* < \infty$ such that*

$$\begin{aligned}\underline{I}_h(r) = I_{\pi^*,h}(r) = \big[\theta^* r - M_{\pi^*,h}(\theta^*)\big] &= \min_{\pi\in\mathbb{P}} \max_{\theta\ge 0} \big[\theta r - M_{\pi,h}(\theta)\big] \\ &= \max_{\theta\ge 0} \min_{\pi\in\mathbb{P}} \big[\theta r - M_{\pi,h}(\theta)\big] = \big[\theta^* r - \overline{M}_h(\theta^*)\big],\end{aligned}$$

where $\overline{M}_h := \log(\overline{m}_h)$ is the worst-case log moment-generating function. □

A geometric interpretation of the extremal property is provided by convexity of the divergence sets: The minimization (25) can be expressed,

$$\underline{I}_h(r) = \inf_{\pi\in\mathbb{P}} \inf_{\mu\in\mathcal{H}\cup\mathcal{H}^1} D(\mu \parallel \pi), \qquad r \in (\overline{r}_h, \overline{h}). \tag{26}$$

Which is equivalently expressed,

$$\underline{I}_h(r) = \sup\{\beta : \mathcal{Q}^+_\beta(\mathbb{P}) \cap \mathcal{H} = \emptyset\}, \qquad r \in (\overline{r}_h, \overline{h}). \tag{27}$$

This follows from the geometry illustrated in Figure 1.

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 1.4 asserts that there exists $\pi^* \in \mathbb{P}$ such that $D(\mu^* \parallel \pi^*) = \beta^*$. The pair of probability distributions $\{\mu^*, \pi^*\}$ solve (26), and π^* is an extremal distribution.

The remainder of the paper is organized as follows. The next section develops two general applications to queueing theory and information theory. Section 3.1 contains a development of theory related to the geometry illustrated in Figure 1. In particular, the geometry of divergence sets is explored, and extremal distributions are characterized. Conclusions are contained in Section 4, along with a description of possible future directions and applications for this research.

Proofs of the major results are contained in the Appendix.

2 Applications

We illustrate the application of the Generalized Bennett's Lemma Theorem 1.2 and the worst-case Sanov bound Theorem 1.4 in two general settings. In Section 2.1 we consider bounds on error exponents arising in the analysis of buffer overflows in queues. Section 2.2 contains application to robust hypothesis testing based on Theorem 1.4. Further discussion on the hypothesis testing problem is included in Section 3.3.1, and [33, 31] contains a more complete development in a completely general setting.

2.1 Buffer overflows in queues

The reflected random walk is a basic model in queueing theory. In the most common application, W_t is interpreted as the total workload in the queue at time k, and evolves according to

$$W_{t+1} = [W_t + X_{t+1}]_+, \qquad W_0 \in \mathbb{R}_+,$$

where $\boldsymbol{X}$ is an i.i.d. sequence. For example, X_i might represent the duration of the ith telephone call to a call center, minus the time elapsed since the previous call. The marginal distribution of $\boldsymbol{X}$ may be complex. To obtain a simpler model, one can estimate the first few moments of the marginal distribution, and then compute the worst case marginal that maximizes some cost criterion subject to these moment constraints.

For example, if just two moments are estimated, and if one maximizes $\langle \pi, f \rangle$ over all π subject to these moment constraints with $f(x) = e^{\theta x}$ for some $\theta > 0$, then we have seen that π^* is a binary by Bennett's Lemma. The following is a simple but useful extension:

Proposition 2.1. *Each of the following expectations,*

$$\mathsf{E}[X^p], \quad \mathsf{E}[e^{\vartheta X}], \quad \mathsf{E}[X^p e^{\vartheta X}],$$

is maximized simultaneously for each $\vartheta > 0$, $p \in \mathbb{Z}_+$, by a fixed binary distribution in each of the following two situations:

(i) *If the first moment $c_1 = \int x\, \pi(dx)$ is specified, then each of these expectations is maximized over all probability distributions on $[-\overline{x}^-, \overline{x}^+]$ with mean c_1 by*

$$\pi^* = p^*\, \delta_{-\overline{x}^-} + (1 - p^*)\, \delta_{\overline{x}^+}\,, \tag{28}$$

where $p^* = (\overline{x}^+ - c_1)/(\overline{x}^+ + \overline{x}^-)$.

(ii) *If two moments are specified,* $c_i = \int x^i \, \pi(dx)$, $i = 1, 2$, *then the optimizer is,*

$$\pi^* = p^* \, \delta_{x_0} + (1 - p^*) \, \delta_{\overline{x}^+} \,, \tag{29}$$

where $x_0 = [\overline{x}^+ - c_1]^{-1}(c_1 \overline{x}^+ - c_2)$ *and* $p^* = [\overline{x}^{+2} + c_2 - 2c_1 \overline{x}^+]^{-1}(\overline{x}^+ - c_1)^2$.

The significance of Proposition 2.1 is that a very simple model can be constructed that captures worst-case behavior.

Assume that the marginal distribution π of $\boldsymbol{X}$ is supported on an interval $[-\overline{x}^-, \overline{x}^+]$ with $\overline{x}^-$ and $\overline{x}^+$ each strictly positive. The mean is assumed negative $\mathsf{E}[X_i] = -d < 0$, and $\mathsf{P}\{X_i > 0\} > 0$. This ensures that $\boldsymbol{W}$ is positive recurrent, and its unique invariant measure has non-trivial support.

Under general conditions, the steady state mean of $\boldsymbol{W}$ is determined by only the first and second moments of $\boldsymbol{X}$ (see for example the version of the Pollaczek-Khintchine formula [30, Prop. 1.1].) However, two distributions with common first and second moments may have very different higher-order statistics, and hence exhibit very different behavior during a rare event such as a buffer overflow.

We consider the worst case behavior of the stationary version of the random walk on the two-sided time interval $\mathbb{Z}$. Denoting the stationary process $\boldsymbol{W}^*$, consider the tail exponent,

$$I_W = -\lim_{b \to \infty} b^{-1} \log\big(\mathsf{P}\{W_0^* \geq b\}\big).$$

The log moment generating function for π, denoted Λ, satisfies $\Lambda'(0) = -d$ and $\Lambda(\theta) \to \infty$ as $\theta \to \infty$. Consequently, there is a unique second zero $\theta_0 > 0$. It is known that that $I_W = \theta_0$, and that the derivative $d^+ = \Lambda'(\theta_0)$ is the most likely slope of $\boldsymbol{W}^*$ prior to 'overflow' [9].

Proposition 2.2. *The exponent* I_W *is minimized, and the slope* $d^+ = \Lambda'(\theta_0)$ *is maximized, over all marginals* π *on* $[-\overline{x}^-, \overline{x}^+]$ *with given first moment* $-d$*, by the values obtained when the marginal distribution of* $\boldsymbol{X}$ *is supported on the two points* $\{-\overline{x}^-, \overline{x}^+\}$*. In particular, for any increment distribution supported on this interval with the given mean* $-d$ *we have,*

$$\lim_{b \to \infty} -b^{-1} \log \mathsf{P}\{W_0^* \geq b\}\big) \leq -\theta_0^\bullet,$$

where $\theta_0^\bullet$ *denotes the positive zero of the worst-case log moment generating function*

$$\Lambda_\bullet(\theta) = \log\Big(\frac{(\overline{x}^+ + d)e^{-\overline{x}^- \theta} + (\overline{x}^- - d)e^{\overline{x}^+ \theta}}{\overline{x}^- + \overline{x}^+}\Big), \qquad \theta \in \mathbb{R}.$$

Proof. Applying Proposition 2.1, we see that the binary distribution maximizes $\Lambda(\theta)$ for $\theta > 0$, and hence the location of the second zero θ_0 is minimized over all $\pi \in \mathbb{P}$. Moreover, we have

$$\Lambda'(\theta_0) = \mathsf{E}[X e^{\vartheta_0 X}],$$

which is also maximized for the same reasons. ☐

We close with a numerical example of the following parameterized form. Fix $d \in (0, 1)$, and let $\kappa \geq 1$ denote the parameter. We choose $\overline{x}^- = 1$ and $\overline{x}^+ = \kappa$, so that the worst-case distribution π^* is determined by $p_\kappa = \pi^*\{-1\} = (1+\kappa)^{-1}(d+\kappa)$ to satisfy the mean constraint.

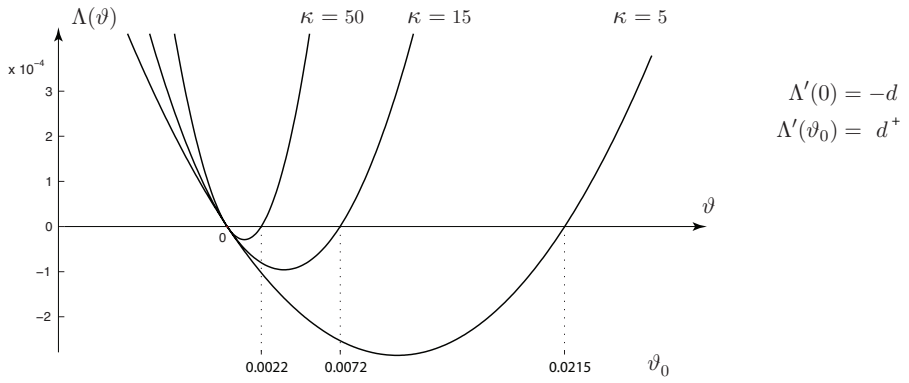

Figure 3: Log moment generating function for three binary distributions supported on $\{-1, \kappa\}$ with $\kappa = 5, 15, 50$, and common mean $\mathsf{E}[X_t] = -d$.

The moment generating function and its derivative are expressed,

$$\lambda(\vartheta) = p_\kappa e^{-\vartheta} + (1 - p_\kappa)e^{\kappa\vartheta}, \qquad \lambda'(\vartheta) = -p_\kappa e^{-\vartheta} + (1 - p_\kappa)\kappa e^{\kappa\vartheta}, \qquad \vartheta \in \mathbb{R}.$$

Through a second-order Taylor series approximation we obtain,

$$1 = \lambda(\vartheta_0) \geq p_\kappa(1 - \vartheta_0) + (1 - p_\kappa)(1 + \kappa\vartheta_0 + \tfrac{1}{2}(\kappa\vartheta_0)^2),$$

which after substituting the definition of p_κ and rearranging terms gives,

$$\limsup_{\kappa\to\infty} \kappa\vartheta_0(\kappa) \leq 2\frac{d}{1-d}.$$

Based on this bound, we obtain through a first-order Taylor series approximation,

$$p_\kappa(1 - \vartheta_0) + (1 - p_\kappa)e^{\kappa\vartheta_0} = 1 + O(\kappa^{-2}),$$

which then implies the limit,

$$\lim_{\kappa\to\infty} \kappa\vartheta_0(\kappa) = B_0,$$

where $B_0 \in (0, 2(1-d)^{-1}d]$ solves the fixed point equation $e^{B_0} = 1 + (1-d)^{-1}B_0$.

As expected, the value of ϑ_0 vanishes as $\kappa \to \infty$. However, the slope $d^+ = \Lambda'(\theta_0) = \lambda'(\vartheta_0)$ is bounded, and converges to the finite limit,

$$d^+(\infty) := \lim_{\kappa\to\infty} d^+(\kappa) = -1 + (1 - d)e^{B_0}.$$

From the fixed point equation and the bound $B_0 \leq 2(1-d)^{-1}d$ this gives,

$$d^+(\infty) = -1 + (1 - d)[1 + (1 - d)^{-1}B_0] = -d + B_0 \leq \Big(\frac{1+d}{1-d}\Big)d.$$

A numerical experiment was conducted with $d = 1/19$.[1] A plot of $\Lambda = \log(\lambda)$ is shown in Figure 3 for $\kappa = 5, 15, 50$. We have $\Lambda'(0) = -d$ for each κ, and $d^+ = \Lambda'(\theta_0)$ is approximately equal to d in each of the three plots.

In conclusion, although the error exponent ϑ_0 is highly sensitive to the parameter κ, the most likely behavior during a rare event is relatively insensitive.

[1] when $\kappa = 1$, this corresponds to a sampled M/M/1 queue with load $\rho = 0.9$.

2.2 Robust hypothesis testing

The most compelling applications of the results developed in this paper concern hypothesis testing, and related topics in information theory.

Consider the following classical hypothesis testing problem. A set of i.i.d. measurements $\{X_1, \dots, X_N\}$ is observed, and one must decide if the observations are generated by one of two given marginal distributions π_0 or π_1, representing the two 'hypotheses' H_0 and H_1. To avoid technicalities we shall assume that the state space X is finite.

For a given $N \geq 1$, suppose that a decision test ϕ_N is constructed based on the finite set of measurements $\{X_1, \dots, X_N\}$, with $\phi_N(X_1, \dots, X_N) = 1$ interpreted as the declaration that hypothesis H_1 is true. The performance of a given test sequence is reflected in the error exponents for the type-II and type-I error probabilities, defined respectively by,

$$I_\phi := -\lim_{N\to\infty} \inf \frac{1}{N} \log \mathsf{P}_{\pi_1}\{\phi_N = 0\}, \qquad J_\phi := -\lim_{N\to\infty} \inf \frac{1}{N} \log \mathsf{P}_{\pi_0}\{\phi_N = 1\},$$

where P_{π_0}, P_{π_1} denote the distributions of $\boldsymbol{X}$ under H_0 and H_1 respectively.

A test is optimal with respect to the (asymptotic) Neyman-Pearson N-P criterion if it maximizes the type-II exponent subject to a constraint on the type-I exponent. Thus, for a given constant $\eta > 0$,

$$\sup_\phi I_\phi \qquad \text{subject to} \qquad J_\phi \geq \eta. \tag{30}$$

The value of the optimization (30) can be expressed as the convex program,

$$\beta^*(\pi_0, \pi_1) = \inf\{D(\mu\|\pi_1) : D(\mu\|\pi_0) \leq \eta\}, \tag{31}$$

and the solution to (31) leads to the well-known log-likelihood ratio test [14, 35].

However, there are many other solutions to (30), and hence many optimal tests. It is shown in [45] that one may restrict to tests of the following form without loss of generality: for a closed set $\mathcal{A} \subseteq \mathcal{M}_1$,

$$\phi_N = \mathbb{I}\{\Gamma_N \in \mathcal{A}^c\} \tag{32}$$

There is a *universal test* of this form that is also optimal, and does not require knowledge of hypothesis H_1: One takes $\mathcal{A} = \mathcal{Q}_\eta^+(\pi_0)$, or equivalently,

$$\phi_N^* = 0 \iff D(\Gamma_N\|\pi_0) \leq \eta. \tag{33}$$

This test is minimal over all tests based on the empirical distributions $\{\Gamma_N\}$. That is, for any test of the form (32) for a closed set $\mathcal{A} \subset \mathcal{M}_1$, if the Neyman-Pearson criterion is satisfied then $\mathcal{Q}_\eta^+(\pi_0) \subset \mathcal{A}$.

In most applications, the statistics under either hypothesis are not completely known apriori. In this case, it may be reasonable to assume that for each i, the measure π_i belongs to a given uncertainty class. A standard approach to designing decision rules in this setting is the min-max "robust" approach, where the goal is to minimize the worst-case performance over the uncertainty classes. Robust detection has been the subject of numerous papers in this setting since the seminal work of Huber and Strassen [18]. A survey of robust hypothesis testing research may be found in [43].

Consider the following robust hypothesis testing problem in which H_i refers to the hypothesis that the marginal distribution belongs to the moment class $\mathbb{P}_i$, $i = 0, 1$. 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 $\pi_0 \in \mathbb{P}_0$:

$$\sup_{\phi} \inf_{\pi_1 \in \mathbb{P}_1} I_\phi^{\pi_1} \qquad \text{subject to} \qquad \inf_{\pi_0 \in \mathbb{P}_0} J_\phi^{\pi_0} \geq \eta. \tag{34}$$

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

We first describe the analog of (33). Let $L_0(\mu) := \inf_{\pi \in \mathbb{P}_0} D(\mu \| \pi)$ denote the minimal relative entropy under H_0. An optimal test can be defined by choosing as an acceptance region for H_0 the sublevel set $\mathcal{Q}_\eta^+(\mathbb{P}_0)$,

$$\phi_N^{*\mathrm{u}} = 0 \iff L_0(\Gamma_N) \leq \eta \tag{35}$$

Proposition 2.3. *Suppose that $\mathbb{P}_0$ satisfies Assumption (A1). Then the test (35) is optimal. Moreover, it is universal in that it maximizes $I_\phi^{\pi_1}$ over all test sequences, regardless of the marginal distribution π_1.*

Proof. The fact that this test achieves the constraint η on the worst-case missed detection exponent follows directly from the fact that $\mathcal{Q}_\eta^+(\mathbb{P}_0) = \{\mu : L_0(\mu) \leq \eta\}$ contains $\mathcal{Q}_\eta^+(\pi_0)$ for any $\pi_0 \in \mathbb{P}_0$.

Conversely, as remarked above, for any $\pi_0 \in \mathbb{P}_0$ the acceptance region $\mathcal{Q}_\eta^+(\pi_0)$ is minimal over all closed subsets of $\mathcal{M}_1$ that give rise to a feasible test. Hence any optimal acceptance region for the min-max problem must contain the union $\bigcup_{\pi_0 \in \mathbb{P}_0} \mathcal{Q}_\eta^+(\pi_0)$, which is precisely $\mathcal{Q}_\eta^+(\mathbb{P}_0)$. ☐

When H_1 is specified via moment constraints then it is possible to construct a simpler optimal test. Although the test itself is not a log-likelihood test, it has a geometric interpretation that is entirely analogous to that given in Hoeffding's result [14]. The value β^* in an optimal test can be expressed,

$$\beta^* = \inf\{\beta : \mathcal{Q}_\eta^+(\mathbb{P}_0) \cap \mathcal{Q}_\beta^+(\mathbb{P}_1) \neq \emptyset\}. \tag{36}$$

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

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

The distribution μ^* has the form $\mu^*(x) = \ell_0(x)\pi_0^*(x)$, where the function ℓ_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 4.

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

Proposition 2.4. *Suppose that $\mathbb{P}_0$ and $\mathbb{P}_1$ each satisfy Assumption (A1). Letting f denote the vector function and c the constant vector that determine $\mathbb{P}_0$, there exists $\lambda \in R(f, c)$ such that with $\ell_0 = \lambda^T f$, the following test is optimal*

$$\phi_N^* = 0 \iff N^{-1} \sum_{t=0}^{N-1} \log(\ell_0(X_t)) \leq \eta. \tag{37}$$

Proof. We appeal to the convex geometry illustrated in Figure 4. 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 $h\colon \mathsf{X} \to \mathbb{R}$ defining a separating hyperplane between the sets $\mathcal{Q}_\eta^+(\mathbb{P}_0)$ and $\mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$, satisfying

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

Theorem 3.2 then implies that there exists $\lambda \in R(f, c)$ such that the function $\log(\lambda^T f)$ also defines a separating hyperplane,

$$\mathcal{Q}_\eta(\mathbb{P}_0) \subset \mathcal{H}_*^0 := \{\mu \in \mathcal{M}_1 : \langle \mu, h\rangle < \eta\}, \quad \mathcal{Q}_{\beta^*}(\mathbb{P}_1) \subset \mathcal{H}_*^1 := \{\mu \in \mathcal{M}_1 : \langle \mu, h\rangle > \eta\}.$$

Moreover $\log(\lambda^T f) \geq h$ everywhere (with equality a.e. μ^*.) This is the vector λ used in (37).

This test is more 'liberal' then the universal test (35), in the sense that it accepts H_0 more frequently, so that the test is feasible. In fact, the worst-case false alarm error exponent is precisely η since $\mu^* \in \partial\mathcal{H}^0$ and $D(\mu^* \| \pi_0^*) = \eta$.

Since $D(\mu^* \| \pi_1^*) = \beta^*$ it follows that the value of (34) can be no greater than this β^*. Conversely, considering the convex geometry shown in Figure 4 we conclude that for any $\pi_1 \in \mathbb{P}_1$, $\mu \in \mathcal{H}_*^0$, we have $D(\mu \| \pi_1) \geq L_1(\mu) \geq L_1(\mu^*) = \beta^*$. Sanov's Theorem then implies that this test achieves this upper bound,

$$\inf_{\pi_1 \in \mathbb{P}_1} I_\phi^{\pi_1} = \inf_{\pi_1 \in \mathbb{P}_1, \mu \in \mathcal{H}_*^0} D(\mu \| \pi_1) = \beta^*.$$

□

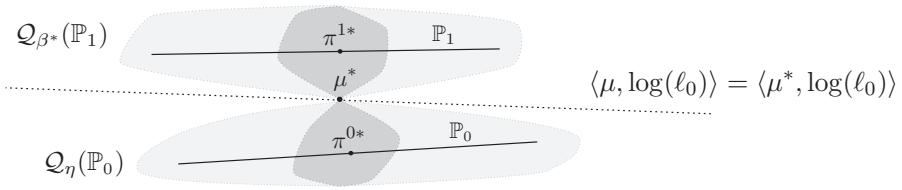

Figure 4: The two-moment worst-case hypothesis testing problem. The uncertainty classes $\mathbb{P}_i$, $i = 0, 1$ are determined by a finite number of linear constraints, and the thickened regions $\mathcal{Q}_\eta(\mathbb{P}_0)$, $\mathcal{Q}_{\beta^*}(\mathbb{P}_1)$ are each convex. The linear threshold test is interpreted as a separating hyperplane between these two convex sets.

In the remainder of this section these tests are illustrated using a simple example.

Variance discrimination Consider the special case in which the two hypotheses are defined by first and second moment constraints. It is assumed that $\mathsf{X} = [-1, 1]$, and the first moments are assumed to be zero, giving

$$\mathbb{P}_0 = \left\{\pi \in \mathcal{M}_1 : \langle \pi, f_1\rangle = 0, \langle \pi, f_2\rangle \leq \sigma_0^2\right\} \quad \mathbb{P}_1 = \left\{\pi \in \mathcal{M}_1 : \langle \pi, f_1\rangle = 0, \langle \pi, f_2\rangle \geq \sigma_1^2\right\} \quad (38)$$

where $f_k(x) \equiv x^k$ for $k = 1, 2$, and $0 < \sigma_0^2 < \sigma_1^2 < 1$ are known bounds on the respective variances.

Note that we have relaxed the assumption that the state space is finite. Also, note that we are considering inequality constraints in this problem. However, we will find the solution is identical to what is obtained using equality constraints.

The solution to the robust hypothesis testing problem is illustrated in Figure 4. The optimal exponent β^* is given by the supremum in (36), and there exists μ^* solving (36) in the sense that $L_0(\mu^*) = \eta$ and $L_1(\mu^*) = \beta^*$. Three tests are considered here:

(i) The universal test that does not depend upon $\mathbb{P}_1$ is expressed,

$$\phi_N^{*u} = 0 \iff N^{-1} \sum_{t=0}^{N-1} \log\bigl(1 + \lambda_1 X_t + \lambda_2 (X_t^2 - \sigma_0^2)\bigr) \le \eta, \tag{39}$$

for every pair λ_1, λ_2 such that the quadratic function $q_0(x) := 1 + \lambda_1 x + \lambda_2(x^2 - \sigma_0^2)$ is non-negative on $[-1, 1]$. Note that $q_0 = \lambda^{\mathrm{T}} f$ with $\lambda = (\lambda_0, \lambda_1, \lambda_2)^{\mathrm{T}} \in R(f, c)$, where we have eliminated λ_0 using the linear constraint $\lambda^{\mathrm{T}} c = \lambda_0 + \lambda_2 \sigma_0^2 = 1$.

(ii) When $\mathbb{P}_1$ is specified by (38) then one can restrict to a single quadratic in an optimal test,

$$\phi_N^* = 0 \iff N^{-1} \sum_{t=0}^{N-1} \log\bigl(1 + \lambda_2 (X_t^2 - \sigma_0^2)\bigr) \le \eta. \tag{40}$$

We show in Proposition 2.5 that, letting $\sigma_\bullet^2 \in (\sigma_0^2, \sigma_1^2)$ denote the variance of μ^*,

$$\lambda_2 = \frac{1}{\sigma_0^2} \frac{\sigma_\bullet^2 - \sigma_0^2}{1 - \sigma_0^2}. \tag{41}$$

(iii) The *naive test* is defined based on the sample-path second moment,

$$\phi_N = 0 \iff N^{-1} \sum_{t=0}^{N-1} X_t^2 \le \tau^2, \tag{42}$$

with $\tau^2 \in (\sigma_0^2, \sigma_1^2)$ a fixed threshold.

Proposition 2.5 asserts that the test (42) is optimal for an appropriate value of τ^2 depending on η. We note that the proof of this result is based on symmetry, and is hence extremely fragile: If for example the state space $\mathsf{X} = [-1, 1]$ is replaced by any non-symmetric interval, or if the zero-mean constraints are modified, then we no longer know if the naive test is optimal.

Proposition 2.5. *Suppose that the two moment classes are defined by (38). Then the test (40) is optimal when λ_2 is given by (41), where the variance parameter $\sigma_\bullet^2$ is the solution to,*

$$\eta = (1 - \sigma_\bullet^2) \log\Bigl(\frac{1 - \sigma_\bullet^2}{1 - \sigma_0^2}\Bigr) + \sigma_\bullet^2 \log\Bigl(\frac{\sigma_\bullet^2}{\sigma_0^2}\Bigr).$$

The naive test (42) using $\tau^2 = \sigma_\bullet^2$ is also optimal, and the three optimal tests satisfy,

$$\phi_N^{*u} = 0 \Longrightarrow \phi_N^* = 0 \Longrightarrow \phi_N = 0.$$

To prove Proposition 2.5 we first demonstrate that the optimizing distributions are symmetric. The conclusion (43) can be equivalently expressed $\mu^* \in \mathcal{Q}_\eta^+(\mathbb{P}_0) \cap \mathcal{Q}_{\beta^*}^+(\mathbb{P}_1)$ since β^* satisfies (31).

Lemma 2.6. *There exist three symmetric distributions π_0^*, π_1^*, μ^* on $[-1,1]$ that solve the robust hypothesis testing problem:*

$$D(\mu^* \| \pi_0^*) = \eta, \quad D(\mu^* \| \pi_1^*) = \beta^*. \tag{43}$$

Proof. For any distribution μ denote by μ° the symmetric distribution defined by $\mu^\circ = \frac{1}{2}(\mu + \tilde{\mu})$, where $\tilde{\mu}(dx) = \mu(-dx)$.

Suppose that $\{\pi_0^*, \pi_1^*, \mu^*\}$ is any triple solving the robust hypothesis testing problem. Then, the triple $\tilde{\pi}_0^*, \tilde{\pi}_1^*, \tilde{\mu}^*$ also forms a solution. Convexity then implies that $L_0(\mu^{*\circ}) \le \eta$ and $L_1(\mu^{*\circ}) \le \beta^*$, and since β^* solves (36) these upper bounds must be achieved. ☐

We henceforth assume that the three distributions are symmetric.

Applying (37) we can construct an optimal test of the form,

$$\phi_N^* = 0 \iff \langle \Gamma_N, \log(q_0) \rangle \le \eta$$

where $q_0(x) = \lambda_0 + \lambda_1 x + \lambda_2 x^2$, with $\lambda \in R(f, c^0)$. Optimality is characterized by the inclusions,

$$\mathcal{Q}_\eta(\mathbb{P}_0) \subset \mathcal{H}_*^0, \qquad \mathcal{Q}_{\beta^*}(\mathbb{P}_1) \subset \mathcal{H}_*^1, \tag{44}$$

where,

$$\mathcal{H}_*^0 := \{\mu : \langle \mu, \log(q_0) \rangle < \eta\}, \qquad \mathcal{H}_*^1 := \{\mu : \langle \mu, \log(q_0) \rangle > \eta\}.$$

Lemma 2.7. *The quadratic q_0 is of the form, for some $\lambda_2 \in \mathbb{R}$,*

$$q_0(x) = 1 + \lambda_2(x^2 - \sigma_0^2), \qquad x \in \mathbb{R}.$$

Proof. Theorem 1.4 implies that μ^* can be expressed,

$$\frac{d\mu^*}{d\pi_0^*} = q_0, \qquad a.e.\ [\pi_0^*].$$

Symmetry of μ^* and π_0^* implies that $\lambda_1 = 0$. Moreover, since μ^* is a probability measure we have $1 = \mu^*(\mathsf{X}) = \pi_0^*(q_0) = \lambda_0 + \lambda_2 \sigma_0^2$, giving $\lambda_0 = 1 - \lambda_2 \sigma_0^2$. ☐

Switching the roles of η and β^* we obtain,

Lemma 2.8. *There exists a quadratic function of the form,*

$$q_1(x) = 1 + \gamma_2(x^2 - \sigma_1^2), \qquad x \in \mathbb{R},$$

with $\gamma_2 \in \mathbb{R}$, such that q_1 is non-negative on $[-1,1]$, and

$$\mathcal{Q}_{\beta^*}(\mathbb{P}_1) \subset \mathcal{H}_{**}^0 := \{\mu : \langle \mu, \log(q_1) \rangle < \beta^*\} \qquad \mathcal{Q}_\eta(\mathbb{P}_0) \subset \mathcal{H}_{**}^1 := \{\mu : \langle \mu, \log(q_0) \rangle > \beta^*\}.$$

☐

Lemma 2.9. *The two probabilities π_0^*, π_1^* have just three points of support $\{-1, 0, 1\}$, and hence the explicit form,*

$$\pi_i^* = \tfrac{1}{2}\sigma_i^2(\delta_1 + \delta_{-1}) + (1 - \sigma_i^2)\delta_0.$$

The support of μ^* is identical, with

$$\mu^* = \tfrac{1}{2}\sigma_\bullet^2(\delta_1 + \delta_{-1}) + (1 - \sigma_\bullet^2)\delta_0.$$

Proof. Lemma 2.8 asserts that $(\log(q_1), \beta^*)$ defines a tangent hyperplane to $\mathcal{Q}_\eta(\mathbb{P}_0)$ passing through μ^*, which we write as

$$\mathcal{Q}_\eta(\mathbb{P}_0) \subset \{\mu : \langle \mu, -\log(q_1)\rangle < -\beta^*\}.$$

We can thus apply Theorem 3.2 to find $\theta_* > 0$ satisfying,

$$\log(q_0(x)) - \eta \geq \theta_*(-\log(q_1(x)) + \beta^*), \qquad x \in [-1, 1],$$

with equality almost everywhere. Writing $y = x^2$ this becomes,

$$\log(\lambda_0 + \lambda_2 y) - \eta \geq \theta_*[-\log(\gamma_0 + \gamma_2 y) + \beta^*], \qquad y \in [0, 1], \tag{45}$$

again with equality almost everywhere. The left hand side of this inequality is a strictly concave function of y, and the right hand side is convex. It follows that equality can hold only at the two points $\{0, 1\}$. That is, for some $p_i \in (0, 1)$,

$$\pi_i^* = \tfrac{1}{2}p_i\delta_1 + \tfrac{1}{2}p_i\delta_{-1} + (1 - p_i)\delta_0.$$

From the variance constraint we obtain $p_i = \sigma_i^2$.

Identical reasoning yields the desired representation for μ^*. □

Proof of Proposition 2.5. To prove the proposition we establish an analog of (44),

$$\begin{aligned} \mathcal{Q}_\eta(\mathbb{P}_0) &\subset \mathcal{H}^0 := \{\mu : \langle \mu, f_2\rangle < \tau^2\}, \\ \mathcal{Q}_{\beta^*}(\mathbb{P}_1) &\subset \mathcal{H}^1 := \{\mu : \langle \mu, f_2\rangle > \tau^2\}. \end{aligned} \tag{46}$$

This entails constructing a linear function $\ell(y)$ satisfying,

$$\log(\lambda_0 + \lambda_2 y) - \eta \geq \ell(y) \geq \theta_*[-\log(\gamma_0 + \gamma_2 y) + \beta^*], \qquad y \in [0, 1]. \tag{47}$$

This is clearly possible, due to the alignment condition (45), and we necessarily have equality at the two endpoints. In particular, $\ell(0) = \log(\lambda_0) - \eta$.

Writing $\ell(x^2) = A + Bf_2(x)$ we obtain,

$$0 = \langle \mu^*, \log(\lambda_0 + \lambda_2 f_2) - \eta\rangle = \langle \mu^*, A + Bf_2\rangle = A + B\tau^2,$$

giving $B = -A^{-1}\tau^{-2}$, and $A = \ell(0) = \log(\lambda_0) - \eta$. These two equations can be combined to give $\ell(x^2) = \theta_\bullet(x^2 - \tau^2)$ with $\theta_\bullet := \tau^{-2}(\eta - \log(\lambda_0))$.

Given this form we can rewrite (47) as follows,

$$\log(q_0(x)) - \eta \geq \theta_\bullet(f_2(x) - \tau^2) \geq \theta_*[-\log(q_1(x)) + \beta^*], \qquad x \in [-1, 1]. \tag{48}$$

To see that this implies (46) we note that the two inequalities in (48) imply the following corresponding inclusions,

$$\mathcal{Q}_\eta(\mathbb{P}_0) \subset \{\mu : \langle \mu, \log(q_0) - \eta \rangle < 0\} \subset \{\mu : \langle \mu, \theta_\bullet(f_2 - \tau^2) \rangle < 0\} = \mathcal{H}^0,$$

$$\mathcal{Q}_{\beta^*}(\mathbb{P}_1) \subset \{\mu : \langle \mu, \log(q_1) \rangle - \beta^* < 0\} \subset \{\mu : \langle \mu, {\theta_*}^{-1}\theta_\bullet(-f_2 + \tau^2) \rangle < 0\} = \mathcal{H}^1.$$

To complete the proof we now obtain the expressions for λ_2 and τ^2. The formula for τ^2 follows from the relative entropy expression,

$$\eta = D(\mu^* \| \pi_0^*) = \langle \mu, \log(d\mu^*/d\pi^*) \rangle,$$

and the expressions for μ^* and π_*^0 given in Lemma 2.9.

The formula for λ_2 also follows from Lemma 2.9, and the expression $\mu^* = q_0\pi_0^*$:

$$\tau^2 = \mu^*(-1) + \mu^*(1) = q_0(-1)\pi_0^*(-1) + q_0(1)\pi_0^*(1) = q_0(1)\sigma_0^2.$$

Substituting the identity $q_0(1) = 1 + \lambda_2(1 - \sigma_0^2)$ and solving for λ_2 we obtain (41). ☐

3 Convexity & Alignment

In this section we develop some basic results required in the proof of Theorem 1.4, and various properties of extremal distributions. Recall that $\mathcal{M}_1$ denotes the space of (Borel) probability measures on on the Borel σ-algebra $\mathcal{B}$, endowed with the weak-topology.

We begin in Section 3.1 with an examination of the functional L, and the associated divergence sets defined in (18). We first establish structure for the relative entropy implying that (16) always has a solution:

Theorem 3.1. *The relative entropy $D(\cdot \,\|\, \cdot)$ is jointly convex and lower semi-continuous on $\mathcal{M}_1 \times \mathcal{M}_1$.*

Proof. This follows from the duality relationship,

$$D(\mu \,\|\, \pi) = \sup_{g \in C(\mathsf{X})} \{\langle \mu, g \rangle - \log\langle \pi, e^g \rangle\}$$

The supremum is achieved by the possibly discontinuous function $g = \log(d\mu/d\pi)$. The function $E(\mu, \pi; g) := \langle \mu, g \rangle - \log\langle \pi, e^g \rangle$ is convex and continuous on $\mathcal{M}_1 \times \mathcal{M}_1$ for each $g \in C(\mathsf{X})$, so that its supremum over g is necessarily convex and lower semi-continuous. ☐

3.1 Convex geometry of divergence sets

A characterization of supporting hyperplanes is provided in the next result. For $h \in C(\mathsf{X})$ and $\pi \in \mathcal{M}_1$, we denote by $\overline{h}_\pi$ the essential supremum of h under π.

Theorem 3.2 is the basis of the robust hypothesis testing algorithms surveyed in Section 2.2.

Theorem 3.2. (Identification of Supporting Hyperplanes) *Suppose that $\mu^* \in \partial\mathcal{Q}_\beta^+(\mathbb{P})$, and that $\mathcal{H}$ is a supporting hyperplane for the divergence set $\mathcal{Q}_\beta^+(\mathbb{P})$ at μ^* in the sense that*

$$\mu^* \in \mathcal{Q}_\beta^+(\mathbb{P}) \cap \mathcal{H}, \quad \textit{and} \quad \mathcal{Q}_\beta(\mathbb{P}) \subset \mathcal{H}^0.$$

It is assumed that $\mathcal{H}$ is expressed as (22) for some $r \in (\overline{r}_h, \overline{h})$ (see (24).) Then,

(i) *For each $\pi^* \in \mathbb{P}$ satisfying $D(\mu^* \parallel \pi^*) = \beta$, there are constants $\theta^* > 0$ and $\lambda \in R(f)$ such that,*

$$\begin{aligned} \theta^*(h - r) &\le \log(\lambda^T f) - \beta && \text{everywhere,} \\ \text{and,} \quad \theta^*(h - r) &= \log(\lambda^T f) - \beta = \log\big(\tfrac{d\mu^*}{d\pi^*}\big) - \beta, && \text{a.e. } [\pi^*] \end{aligned} \tag{49}$$

(ii) *Conversely, if $\{\mu^*, \pi^*, \theta^*, \lambda\}$ satisfy (49) with $\pi^* \in \mathbb{P}$, then $\mathcal{H}$ supports $\mathcal{Q}_\beta^+(\mathbb{P})$, with $\mathcal{Q}_\beta(\mathbb{P}) \subset \mathcal{H}^0$.* □

The theorem has several important corollaries described here and in the next subsection. The following result shows that given any supporting hyperplane for $\mathcal{Q}_\beta(\mathbb{P})$ passing through $\mu \in \partial\mathcal{Q}_\beta(\mathbb{P})$, one can construct a supporting hyperplane $\widehat{\mathcal{H}}$ that has a special form, and also passes through μ.

Corollary 3.1. *Let $\mathbb{P}$ be a moment class that satisfies Assumption (A1). For a given $r \in (\overline{r}_h, \overline{h})$ define $\beta = \underline{I}_h(r)$, so that the set $\mathcal{H}$ defined in (22) is a supporting hyperplane for $\mathcal{Q}_\beta^+(\mathbb{P})$, with $\mathcal{Q}_\beta(\mathbb{P}) \subset \mathcal{H}^1$. Then $\hat{h} = \log(\lambda^T f)$ is a continuous function on X, where $\lambda \in R(f)$ is given in Theorem 3.2. Moreover, the set*

$$\widehat{\mathcal{H}} := \{\mu \in \mathcal{M}_1 : \langle \mu, \hat{h} \rangle = \beta\}$$

is also a supporting hyperplane for $\mathcal{Q}_\beta^+(\mathbb{P})$, with

$$\mathcal{Q}_\beta(\mathbb{P}) \subset \widehat{\mathcal{H}}^0 := \{\mu \in \mathcal{M}_1 : \langle \mu, \hat{h} \rangle < \beta\};$$

$$\mathcal{H}^1 \subset \widehat{\mathcal{H}}^1 := \{\mu \in \mathcal{M}_1 : \langle \mu, \hat{h} \rangle > \beta\}.$$

Proof. Since f is a vector of continuous functions, to establish continuity of $\hat{h}$ it is sufficient to prove that $\lambda^T f(x) > 0$ for all $x \in \mathsf{X}$. Since $\mathcal{H}$ supports $\mathcal{Q}_\beta^+(\mathbb{P})$, it must satisfy the necessary conditions of Theorem 3.2 for some $\theta^* > 0$ and $\lambda \in R(f)$. From Theorem 3.2, we know that $\hat{h}$ is bounded below by h, which is bounded. Therefore we do have $\lambda^T f > 0$ on X.

With $\hat{\theta}^* := 1$, it is clear that $\{\hat{h}, \mu^*, \pi^*, \hat{\theta}^*, \lambda\}$ satisfy the sufficient conditions of Theorem 3.2. Thus the hyperplane $\widehat{\mathcal{H}}$ forms a supporting hyperplane for $\mathcal{Q}_\beta^+(\mathbb{P})$ with $\mathcal{Q}_\beta(\mathbb{P}) \subset \{\mu \in \mathcal{M}_1 : \langle \mu, \hat{h} \rangle < \beta\}$. Moreover, since $\theta^*(h - r) \le \hat{h} - \beta$ everywhere, we have

$$\mathcal{H}^1 = \{\mu \in \mathcal{M}_1 : \langle \mu, h - r \rangle > 0\} \subset \{\mu \in \mathcal{M}_1 : \langle \mu, \hat{h} - \beta \rangle > 0\} = \widehat{\mathcal{H}}^1$$

□

We now provide illustrations of the alignment conditions (49) through numerical examples. In each example below the state space is taken to be the unit interval $\mathsf{X} = [0, 1]$. Moment classes are defined using the polynomials $f_i(x) = x^i$, $x \in \mathsf{X}$, with c defined consistently with the uniform distribution ν on $[0, 1]$:

$$c_i = \langle \nu, f_i \rangle = \int_0^1 f_i(x)\, dx = 1/(i+1)^{-1}, \qquad 1 \le i \le n. \tag{50}$$

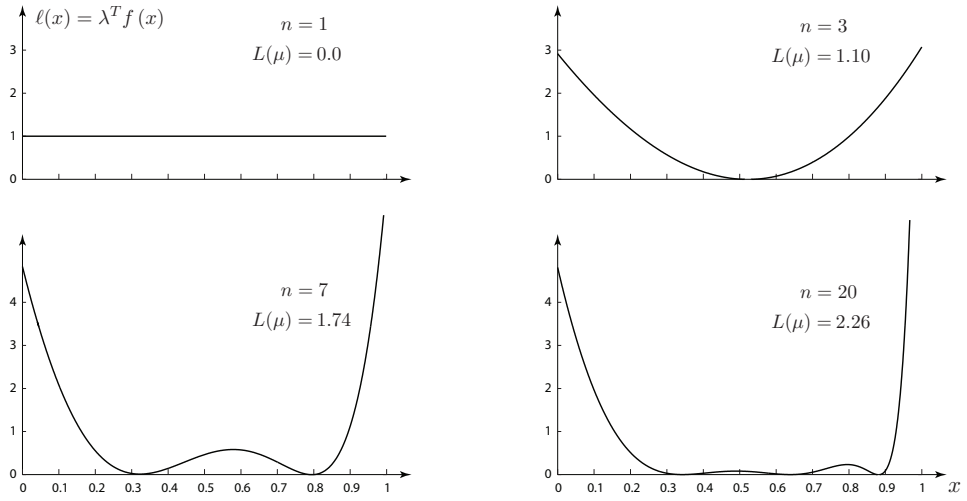

Figure 5: Computation of $L(\mu)$ with μ Bernoulli.

Case I: μ Bernoulli If μ is the symmetric Bernoulli distribution supported on $\{0, 1\}$ then $D(\mu\|\nu) = \infty$. Shown in Figure 5 are results from numerical calculation of L for $n = 1, 3, 7$ and 20. When $n = 1$ we have $c = (1, 0.5)^T$, so that $\mu \in \mathbb{P}$, and hence $L(\mu) = 0$. For each $n \geq 2$ we have $\mu \notin \mathbb{P}$ and, as shown in the figure for three values of n, the worst-case divergence $L(\mu)$ is strictly positive. Also shown in Figure 5 is the nth order polynomial $\lambda^T f$ in each case. From Theorem 1.4 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,

$$\operatorname{supp}(\pi^*) \subset \big\{ \textit{roots of } \lambda^T f \big\} \cup \big\{ \operatorname{supp}(\mu) = \{0, 1\} \big\}$$

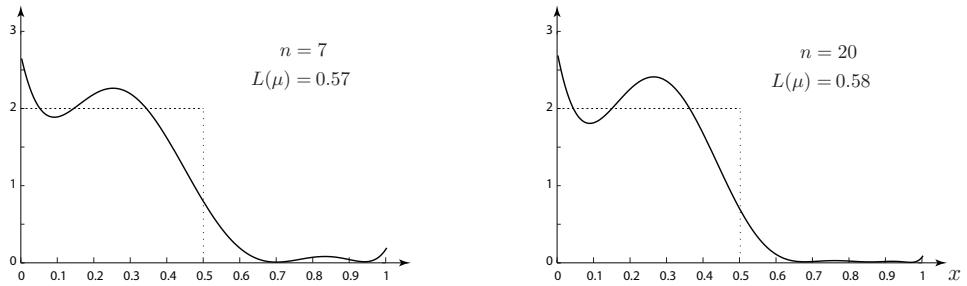

Figure 6: Computation of $L(\mu)$ with μ uniform on $[0, 0.5]$.

Case II: μ uniform If μ is uniform on $[0, 0.5]$ then $D(\mu\|\nu) = \log(2) \approx 0.69$. Shown in Figure 6 is the nth order polynomial $\lambda^T f$ for $n = 7$ and $n = 20$ with $\lambda \in R(f, c)$ and $L(\mu) = \langle \mu, \log(\lambda^T f) \rangle$. In each case this function roughly approximates the density of μ with respect to ν, which is $2\mathbb{I}_{[0,0.5]}$.

3.2 Extremal distributions

We have seen that $\underline{I}_h$ can be computed in (25) by first constructing the worst-case rate-function $L\colon \mathcal{M}_1 \to \mathbb{R}_+$, and then applying the contraction principle. Here we consider the alternate representation of $\underline{I}_h$ obtained from the worst-case log moment generating function, leading to a proof of Theorem 1.5.

Moreover, on analysing general infinite-dimensional linear programs of the form (7) we demonstrate that, without any loss of generality, extremal distributions can be assumed discrete, with no more than $n+2$ points of support.

The *dual* of the general linear program (2) is expressed as,

$$\min\{\lambda^{\mathrm{T}} c : \lambda \in \mathbb{R}^{n+1} \text{ s.t. } \lambda^{\mathrm{T}} f \geq g\}, \tag{51}$$

where $\lambda^{\mathrm{T}} f \geq g$ means $\lambda^{\mathrm{T}} f(x) \geq g(x)$ for all $x \in \mathsf{X}$. It is known that there is no duality gap under Assumption (A1), i.e., the value of the primal (2) is equal to the value of the dual (51). The following result is required in an analysis of the linear program (7). A proof is contained in [22, 39].

Theorem 3.3. (Lack of duality gap) *Under Assumption (A1), for any $g \in C(\mathsf{X})$,*

$$\max\{\langle \pi, g\rangle : \pi \in \mathbb{P}\} = \min\{\lambda^{\mathrm{T}} c : \lambda \in \mathbb{R}^{n+1} \ \textit{s.t.}\ \lambda^{\mathrm{T}} f \geq g\}. \tag{52}$$

Any distribution π^* and vector λ^* optimizing the respective linear programs (2) and (51) are together called a dual pair. A necessary and sufficient condition for a given pair (π, λ) to form a dual pair is the alignment condition*,*

$$\lambda^{\mathrm{T}} f - g \geq 0 \quad \textit{and} \quad \langle \pi, \lambda^{\mathrm{T}} f - g\rangle = 0. \tag{53}$$

□

In this section we focus on the specific linear program that defines $\overline{m}_h$ in (7). Theorem 1.5 provides the following alternate expression for the worst-case rate-function defined in (25):

$$\underline{I}_h(r) = \max_{\theta \geq 0} \big[\theta r - \overline{M}_h(\theta)\big], \qquad r \in (\overline{r}_h, \overline{h}). \tag{54}$$

Recall that in the setting of Theorem 1.1 we have

$$\underline{I}_h(r) = \max_{\theta \geq 0} \big[\theta r - M_{\pi^*,h}(\theta)\big], \qquad r > c_1,$$

with π^* independent of r. When the assumptions of Theorem 1.1 are relaxed then this uniform optimality no longer holds. Shown in Figure 7 are numerical results obtained using $h(x) = 2(x-1) - x\sin(2\pi x)$, $x \in [0,1]$. The moment class $\mathbb{P}$ was defined using polynomials (see (6)), with $n = 2$ and $n = 5$. The worst-case log moment-generating function is plotted for $\theta \in [0, 5]$. Also, for $\theta = \theta^* = 2.5$, the distribution π^* that maximizes (7) was computed, and the figure shows the log moment-generating function $M_{\pi^*,h}$. As required by Theorem 1.5, the functions $\overline{M}_h$ and $M_{\pi^*,h}$ coincide at $\theta = \theta^*$. However, the inequality $\overline{M}_h(\theta) \geq M_{\pi^*,h}(\theta)$ is *strict* for $\theta \in (0, \theta^*)$ and $\theta \in (\theta^*, 5]$.

Proposition 3.4 provides justification for the correspondences illustrated in Figure 2.

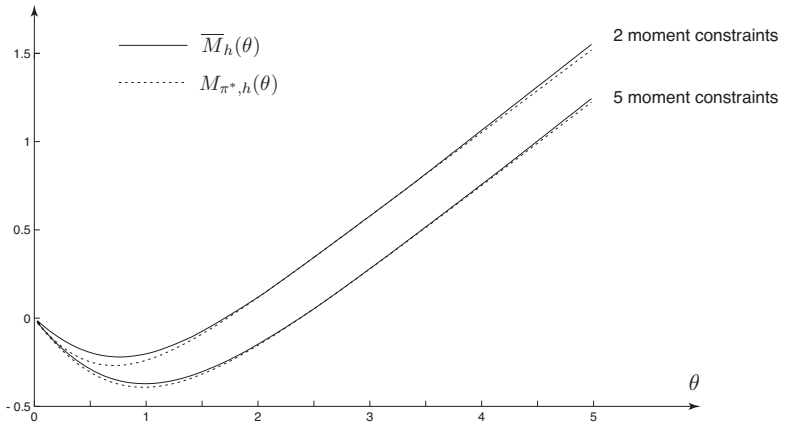

Figure 7: The worst-case log moment-generating function $\overline{M}_h$

Proposition 3.4. *The worst-case log moment-generating function satisfies,*

(i) $\dfrac{d^+}{d\theta}\,\overline{M}_h\,(0) = \overline{r}_h$, *where the 'plus' denotes the right derivative;*

(ii) *The constant* $\overline{r}_h$ *can be expressed as the solution to the linear program,*

$$\overline{r}_h = \max_{\pi \in \mathbb{P}} \langle \pi, h \rangle;$$

(iii) $\lim\limits_{\theta \to \infty} \dfrac{d^+}{d\theta}\,\overline{M}_h\,(\theta) = \overline{h}.$ □

The maximization (7) is a linear program, and is therefore achieved at the extreme points of the set $\mathbb{P}$. These extreme points correspond to discrete measures, and thus (7) suggests (though it does not *imply*), that π^* is a discrete measure. The following theorem establishes that without loss of generality, an extremal distribution can be assumed discrete.

Theorem 3.5. (Discrete extremal distributions) *Under Assumption (A1) suppose that* h *is continuous, and that* $r \in (\overline{r}_h, \overline{h})$. *Then there exists a probability distribution* $\pi^\circ \in \mathbb{P}$ *that is discrete, with no more than* $n+2$ *points of support, and is also* $(r, h, +)$*-extremal.*

Proof. Theorem 1.5 implies that an $(r, h, +)$-extremal distribution $\pi^* \in \mathbb{P}$ is also a solution to the infinite dimensional linear program (7) for some $\theta^* \geq 0$, and that $I_{\pi^*,h}(r) = \underline{I}(r)$. Since $M_{\pi^*,h}$ is analytic and strictly convex on $\mathbb{R}$, the maximality property (5) implies that

$$\frac{d}{d\theta} M_{\pi^*,h}\,(\theta^*) = r.$$

To prove the theorem we construct a discrete probability distribution $\pi^\circ \in \mathbb{P}$ that satisfies $M_{\pi^\circ,h}(\theta^*) = M_{\pi^*,h}(\theta^*) = \overline{M}_h(\theta^*)$, and also the consistent derivative constraint,

$$\frac{d}{d\theta} M_{\pi^\circ,h}\,(\theta^*) = r. \tag{55}$$

It will then follow that $I_{\pi^\circ,h}(r) = I_{\pi^*,h}(r) = \underline{I}(r)$, which is the desired conclusion.

The derivative can be computed for any $\pi \in \mathcal{M}_1$ as follows,

$$\frac{d}{d\theta} M_{\pi,h}(\theta) = \frac{\langle \pi, he^{\theta h} \rangle}{\langle \pi, e^{\theta h} \rangle}, \qquad \theta \in \mathbb{R}.$$

Consequently, equation (55) is expressed as the equality constraint $\langle \pi, he^{\theta^* h} \rangle = r \langle \pi, e^{\theta^* h} \rangle$.

Consider then the linear program,

$$\begin{aligned} \max \langle \pi, \exp(\theta^* h) \rangle \qquad & \text{s.t. } \langle \pi, f_i \rangle = c_i, \qquad i = 0, \dots, n \\ & \langle \pi, he^{\theta^* h} - re^{\theta^* h} \rangle = 0. \end{aligned}$$

This is feasible since the $(h, r, +)$-extremal distribution π^* is one solution. Without loss of generality, we may search among extreme points of the constraint set in this linear program. Since there are $n + 2$ linear contraints, an extreme point may have no more than $n + 2$ points of support. □

3.3 Algorithms

From Theorem 1.4 it follows that $L(\mu)$ can be expressed as the maximum,

$$L(\mu) = \max_{\lambda \in R(f,c)} \langle \mu, \log \lambda^T f \rangle, \qquad \mu \in \mathcal{M}_1, \tag{56}$$

where $R(f, c)$ denotes the convex set,

$$R(f, c) := R(f) \cap \{\lambda : \lambda^T c = 1\}. \tag{57}$$

We thus arrive at a concave program that can be solved in various different ways.

Algorithmic methods for optimization over probability distributions are developed in several recent papers. The multi-dimensional polynomial moment problem is considered in [36], and [6, 16] consider algorithms for computation of mutual information.

We survey several approaches here since one may have to experiment using different techniques when $\mathbb{P}$ is complex. Throughout it is assumed that X is a compact subset of Euclidean space.

3.3.1 Nonlinear programming

Interior point methods The first-order condition for an interior optimizer of (56) is expressed

$$\xi^* := \langle \mu, h^* - c \rangle = 0, \tag{58}$$

where $h := (\lambda^{*T} f)^{-1} f$. When μ has full support then the condition (58) characterizes an optimizer. Hence (56) can be solved using standard interior-point optimization algorithms such as the conjugate gradient algorithm (e.g. [28, 2, 4].)

The *logarithmic barrier method* introduces a 'cost' for approaching the boundary of $R(f, c)$. One version can be expressed as the concave program,

$$\sup\{(1 - \epsilon)\langle \mu, \log \lambda^T f \rangle + \epsilon \langle \nu, \log \lambda^T f \rangle : \lambda \in R(f, c)\}, \tag{59}$$

where ν is any fixed distribution with full support, and $\varepsilon \in (0, 1)$ is the barrier parameter. The solution of the relaxation (59) is precisely $L((1 - \epsilon)\mu + \epsilon\nu)$. The relaxation can again be solved using standard methods.

Projected gradient methods In the case of a boundary optimizer for the original program (56), the vector ξ^* is not necessary zero, but rather satisfies the following inequality constraints:

Proposition 3.6. *The vector $\lambda^\circ \in R(f,c)$ optimizes (56) if and only if the gradient $\xi^\circ := \langle \mu, f({\lambda^\circ}^T f)^{-1}\rangle - c$ satisfies the family of inequalities,*

$$v^T \xi^\circ \le 0, \qquad v \in \mathcal{T}_{\lambda^\circ}(f,c), \tag{60}$$

where $\mathcal{T}_{\lambda^\circ}(f,c) \subset \mathbb{R}^{N+1}$ denotes the set of feasible directions *satisfying $v^T c = 0$, and for some $r_0 > 0$,*

$$(\lambda^\circ + r v^T) f(x) \ge 0 \quad 0 \le r \le r_0.$$

□

The following algorithm was successfully used to compute $L(\mu)$ in the examples illustrated in Figures 5 and 6. For any $\lambda \in R(f,c)$ we define Π_λ to be the projection from $\mathbb{R}^{N+1}$ to $\mathcal{T}_\lambda(f,c)$. Given any $\lambda^t \in R(f,c)$, an ascent direction $v \in \mathcal{T}_{\lambda^t}(f,c)$ is computed satisfying $v^T \xi^t > 0$, where $\xi^t := \langle \mu, h_t\rangle - c$ with $h_t = ({\lambda^t}^T f)^{-1} f$. For example, one can take

$$v = \Pi_{\lambda^t}\{\xi^t\}, \quad \textit{or} \quad v = \Pi_{\lambda^t}\{\Sigma_t \xi^t\}$$

where $\Sigma_t^{-1} := \langle \mu, h_t h_t^T\rangle$ is the negative of the Hessian of $\langle \mu, \log \lambda^T f\rangle$, evaluated at λ^t.

Once a vector $v = v^t$ is selected, the next vector is computed using line-maximization: $\lambda^{t+1} = \lambda^t + r_t v^t$ where,

$$r_t = \arg\max_{r \ge 0} \{\big\langle \mu, \log\big((\lambda^t + r v^t)^T f\big)\big\rangle : \lambda^t + r v^t \in R(f,c)\}. \tag{61}$$

Constraint relaxation To simplify either of the algorithms described above one can simplify the state space. For example, the line search (61) is easily computed when X is finite. One can then introduce new points as the estimate of λ^* is refined.

Suppose that $\mathsf{X}^t = \{x^1, \dots, x^t\} \subset \mathsf{X}$ is a finite set, and suppose that λ^{*t} is the optimal solution to (56) for the reduced state space. We have $\lambda^{*t} \in R_n(f,c)$, meaning that $c^T \lambda^{*t} = 1$, and ${\lambda^{*t}}^T f$ is non-negative on X^t. Since this amounts to a relaxation of (56) we have $L(\mu) \le L_n(\mu) = \langle \mu, \log {\lambda^{*t}}^T f\rangle$. If $\lambda^{*t} \in R(f,c)$ then this inequality is achieved, so that $\lambda^{*t} = \lambda^*$. Otherwise, we set $\mathsf{X}^{t+1} = \mathsf{X}^t \cup \{x^{t+1}\}$ with,

$$x^{t+1} = \arg\min_{x \in \mathsf{X}} \{{\lambda^{*t}}^T f(x)\}.$$

This is similar to the steepest ascent algorithm introduced in [16], which is shown to be convergent. The same arguments can be used to show that $\lambda^{*t} \to \lambda^*$ as $t \to \infty$.

3.3.2 On-line algorithms

An apparent difficulty with the Neyman-Pearson test defined in (37) is that the nonlinear program (56) must be solved to compute $L(\Gamma_N)$ as each new sample is obtained. To make this test practical we require a recursive formula for $L(\Gamma_N)$.

Based on the unconstrained first order condition (58), or the refined first order condition described in Proposition 3.6, we arrive at the Robbins-Monro, stochastic-approximation recursion,

$$\lambda^{t+1} = \lambda^t + \gamma_t \Pi_{\lambda^t}\{h_{t+1}\}, \qquad n \geq 0, \tag{62}$$

where $h_t := (\lambda^{t\,T} f(X_t))^{-1} f(X_t)$, and $\{\gamma_t\}$ is a positive gain sequence satisfying standard assumptions [11].

A refinement of (62) is obtained by mimicking the constrained Newton-Raphson recursion, $\lambda^{t+1} = \lambda^t + \gamma_t \Pi_{\lambda^t}\{\Sigma^t h_t\}$ where here Σ_t^{-1} is an approximation to $\langle \mu, h_t h_t^T \rangle$. Given the sequence of estimates,

$$\Sigma_t^{-1} = \Sigma_t^{-1} - \gamma_t(\Sigma_t^{-1} - h_t h_t^T),$$

we arrive at the following algorithm by applying the Matrix Inversion Lemma [11]:

$$\lambda^{t+1} = \lambda^t + \gamma_t \Pi_{\lambda^t}\{\Sigma_t h_{t+1}\} \tag{63a}$$

$$\Sigma_{t+1} = \Sigma_t - \gamma_t \frac{\Sigma_t + \Sigma_t h_t h_t^T \Sigma_t}{1 + h_t^T \Sigma_t h_t}, \tag{63b}$$

where again $h_t = (\lambda^{t\,T} f(X_t))^{-1} f(X_t)$.

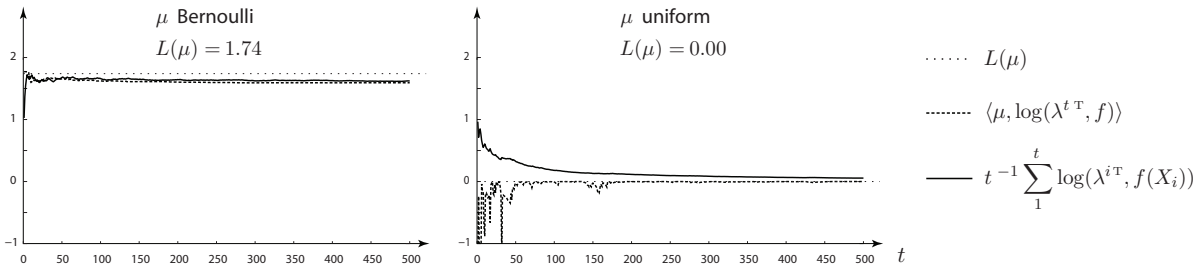

Figure 8: The plot shows $\langle \mu, \log \lambda^{t\,T} f \rangle$ and the running sample path average of $\{\log(\lambda^{i\,T} f(X_i))\}$, where the sequence of vectors $\{\lambda^t\}$ is obtained using the algorithm (62). In the plot shown at left the observations $\{X_i\}$ were Bernoulli, and in the plot at right the observations were uniformly distributed.

To illustrate the application of these stochastic approximation algorithms, consider the computation of $L(\mu)$ when $\mathsf{X} = [0, 1]$, $f(x) = (1, x, \ldots, x^n)^T$, and $\mathbb{P}$ is consistent with the uniform distribution (see (50).) We consider two cases: in the first, the marginal distribution μ of $\boldsymbol{X}$ was Bernoulli as in the results shown in Figure 5. In the second μ was uniform on $[0, 1]$. Seven moment constraints were specified, so that $L(\mu) \approx 1.74$ with μ Bernoulli by the results shown in Figure 5 with $n = 7$, while $L(\mu) = 0$ when μ is uniform since $\mu \in \mathbb{P}$ for any n. The step size was taken to be $\gamma_t = t^{-1}$.

Figure 8 shows results obtained when implementing (62). The two plots are (i) the running sample path average of $\{\log(\lambda^{i\,T} f(X_i))\}$, and (ii) $\langle \mu, \log \lambda^{t\,T} f \rangle$ for $t \geq 1$. In the plot shown at left the marginal distribution of the observations $\boldsymbol{X}$ were Bernoulli, and in the plot at right the marginal was uniform on $[0, 1]$. The algorithm is slow to converge to the precise value of $L(\mu)$, but it distinguishes the observations from $\mathbb{P}$ rapidly.

Similar conclusions were obtained when μ was uniform on $[0, 0.5]$, or a mixture of a Bernoulli and a uniform distribution: The rate of convergence was slow, but the sample path average

of $\{\log(\lambda^{i\,T} f(X_i))\}$ remained strictly positive, maintaining 80% of its final value after a short transient period.

The performance of the more complex algorithm (63a, 63b) was similar. Also, as can be expected, the convergence became slower for larger values of n since λ^t is $(n+1)$-dimensional.

4 Conclusions & Future Directions

We have established explicit formulae for the worst-case large deviations rate-functions appearing in Sanov's Theorem and Chernoff's bound. The geometric structure of the divergence set $\mathcal{Q}_\beta^+(\mathbb{P})$ plays a central role in interpreting the results, and is a valuable tool in analysis.

Potential directions for future research include,

(a) Structure of the worst-case, one-dimensional rate-function defined in (10) deserves further consideration. In particular, under what conditions outside of the polynomial case is π^* independent of $r > 0$?

(b) We have not dealt with methods for selecting the functions $\{f_i\}$ in a particular application.

(c) Extensions of the results here to Markov processes may be possible by applying recent results on exact large deviations [24, 25], and on finite-n bounds [23]. It appears that the formulation of an appropriate moment class is non-trivial for Markov models.

(d) Results from [31] and this paper provided inspiration for the research described in [16, 15, 17]. The discrete nature of extremal distributions provided motivation a new class of algorithms for the computation of efficient channel codes based on optimal discrete input distributions. It is likely that a worst-case approach to channel modeling based on moment classes will lead to simple coding approaches, and easily implemented decoding algorithms.

We are convinced that the theory of extremal distributions will have significant impact in many other areas that involve statistical modeling and prediction.

Acknowledgements Thanks to Professors I. Kontoyiannis and O. Zeitouni for advice on the hypothesis testing literature, and insightful comments.

A Appendix

We collect here proofs of the main results, and several complementary results. We begin with a bound on L, and a description of the domain of $\underline{I}_h$.

Lemma A.1. *Let $\mathbb{P}$ be a moment class that satisfies Assumption (A1). Then,*

(i) *The functional $L\colon \mathcal{M}_1 \to \mathbb{R}_+$ is uniformly bounded:*

$$\sup_{\mu\in\mathcal{M}_1} L(\mu) < \infty.$$

(ii) *The function $\underline{I}_h$ is uniformly bounded on $[\underline{h},\overline{h}]$:*

$$\sup_{r\in[\underline{h},\overline{h}]} \underline{I}_h(r) < \infty.$$

Proof. Define $c_\mu := \langle \mu, f\rangle$ for $\mu \in \mathcal{M}$. This satisfies the uniform bound $\|c_\mu\| \le \max_{x\in\mathsf{X}} \|f(x)\|$ for $\mu \in \mathcal{M}$.

Under (A1) the vector c lies in the interior of Δ. This assumption and the uniform bound on c_μ implies that there exists $\epsilon > 0$ independent of μ such that,

$$\frac{c - \epsilon\, c_\mu}{1-\epsilon} \;\in\; \Delta, \qquad \mu \in \mathcal{M}_1.$$

By the definition of Δ, this means that there exits $\pi \in \mathcal{M}_1$ such that $\langle \pi, f\rangle = \frac{c-\epsilon c_\mu}{1-\epsilon}$. Let $\pi^\epsilon := \epsilon\mu + (1-\epsilon)\pi$. Then $\pi^\epsilon \in \mathbb{P}$, and

$$D(\mu \parallel \pi^\epsilon) \le -\log\epsilon < \infty$$

Thus $L(\mu) = \inf_{\pi\in\mathbb{P}} D(\mu \parallel \pi) \le |\log\epsilon|$ for all $\mu \in \mathcal{M}_1$, and this establishes (i).

Since h is continuous we can find $\overline{x}, \underline{x} \in \mathsf{X}$ such that $h(\overline{x}) = \overline{h}$ and $h(\underline{x}) = \underline{h}$. Moreover, exactly as in the construction of π^ϵ above, we can construct $\pi \in \mathbb{P}$ such that,

$$\overline{p} := \pi\{\overline{x}\} > 0, \quad \underline{p} := \pi\{\underline{x}\} > 0.$$

We then have,

$$M_{\pi,h}(\theta) \ge \log(\underline{p}e^{\underline{h}\theta} + \overline{p}e^{\overline{h}\theta}), \qquad \theta \in \mathbb{R}. \tag{64}$$

Consequently, for $r \in [\underline{h},\overline{h}]$,

$$\begin{aligned}
\sup_{\theta\ge 0}[\theta r - M_{\pi,h}(\theta)] &\le \sup_{\theta\ge 0}\big[\theta(r-\overline{h}) - \log(\overline{p})\big] \le |\log(\overline{p})|,\\
\sup_{\theta\le 0}[\theta r - M_{\pi,h}(\theta)] &\le \sup_{\theta\le 0}\big[\theta(r-\underline{h}) - \log(\underline{p})\big] \le |\log(\underline{p})|.
\end{aligned}$$

This shows that $I_{\pi,h}$ is bounded on $[\underline{h},\overline{h}]$. Minimality of $\underline{I}_h$ completes the proof. ∎

The following result allows us to restrict to a compact domain in the maximization (20).

Lemma A.2. *Suppose that Assumption (A1) holds. Then, the set $R(f,c) \subset \mathbb{R}^{n+1}$ defined in (57) is convex and compact.*

Proof. It is obvious that $R(f,c)$ is closed and convex.

To complete the proof we show that $R(f,c)$ is bounded. Let e^i denote the ith standard basis vector in $\mathbb{R}^n$. Since $(c_1,\dots,c_n)^{\scriptscriptstyle T}$ lies in the interior of the set Δ by Assumption (A1), there exists $\epsilon > 0$ such that $\{(c_1,c_2,..,c_n)^{\scriptscriptstyle T} \pm \epsilon\, e^i : i = 1,\dots,n\} \subset \Delta$

Now from the definition of $R(f)$ it follows that

$$\begin{aligned} R(f) &= \{\lambda \in \mathbb{R}^{n+1} : \lambda^{\scriptscriptstyle T}\langle \pi, f\rangle \ge 0, \text{ for each } \pi \in \mathcal{M}_1\} \\ &= \{\lambda \in \mathbb{R}^{n+1} : \lambda^{\scriptscriptstyle T}(1,x) \ge 0 \text{ for each } x \in \Delta\}, \end{aligned}$$

where $(1,x)^{\scriptscriptstyle T} := (1,x_1,...,x_n)^{\scriptscriptstyle T}$. Using the fact that $(c_1,....,c_n)^{\scriptscriptstyle T} + \epsilon\, e^i \in \Delta$ we conclude

$$\lambda^{\scriptscriptstyle T}(1,(c_1,\dots,c_n)^{\scriptscriptstyle T} + \epsilon\, e^i) \;\ge\; 0,$$

and since $\lambda^{\scriptscriptstyle T} c = 1$ it then follows that $1+\epsilon\lambda_i \ge 0$. Similar reasoning gives the bound $1-\epsilon\lambda_i \ge 0$. Repeating this argument for each $i = 1,\dots,n$, we can infer that

$$\lambda_i \in [-\epsilon^{-1},\, +\epsilon^{-1}],\;\; i = 1,\dots,n.$$

Since $\lambda^{\scriptscriptstyle T} c = 1$ and $c_0 = 1$, the above bounds imply upper and lower bounds on λ_0 as well. Hence $R(f,c)$ is closed and bounded, hence compact, as claimed. ◻

The following version of Cramér's Theorem is used repeatedly below.

Theorem A.3. *Suppose that $\boldsymbol{X}$ is i.i.d. with one dimensional distribution π on $\mathcal{B}$. Fix $h \in C(\mathsf{X})$, and $r \in [\langle \pi, h\rangle, \overline{h}_\pi)$ where $\overline{h}_\pi$ denotes the essential supremum of h. Then,*

(i) *The Chernoff bound is asymptotically tight:*

$$\lim_{N\to\infty} \frac{1}{N} \log\Big(\mathsf{P}\Big[\frac{1}{N}\sum_{i=1}^N h(X_i) \ge r\Big]\Big) = \lim_{N\to\infty} \frac{1}{N} \log\Big(\mathsf{P}\Big[\frac{1}{N}\sum_{i=1}^N h(X_i) > r\Big]\Big) = -I_{\pi,h}(r).$$

(ii) *The one-dimensional rate-function has the following representations,*

$$\begin{aligned} I_{\pi,h}(r) &= \sup_{\theta\ge 0}\{\theta r - M_{\pi,h}(\theta)\} \\ &= \inf\{D(\mu \parallel \pi) : \mu \;\; s.t.\; \langle \mu, h\rangle \ge r\} \end{aligned}$$

(iii) *The infimum over μ and the supremum over θ in (ii) are uniquely achieved by some μ^*, θ^* satisfying*

$$\frac{d\mu^*}{d\pi} = \frac{\exp(\theta^* h)}{\langle \pi, \exp(\theta^* h)\rangle}, \qquad \textit{and} \qquad \langle \mu^*, h\rangle = r.$$

Proof. Parts (i) and (ii) follow from Theorem 1.3 and the Contraction Principle.

Part (iii) follows from [40, Theorem 1.5]. ◻

Define the functional $K\colon \mathcal{M} \to \mathbb{R}$ by,

$$K(\pi) := \inf\{D(\mu \parallel \pi)\colon \mu \in \mathcal{H}^1 \cup \mathcal{H}\}, \qquad \pi \in \mathbb{P}. \tag{65}$$

The following result is an application of Cramér's Theorem:

Lemma A.4. *For each $\pi \in \mathcal{M}$ we have,*

$$K(\pi) = \sup_{\theta \geq 0} \big[\theta r - M_{\pi,h}(\theta)\big]. \tag{66}$$

Proof. This follows directly from Theorem A.3 when $\pi \in \mathcal{M}_1$. Moreover, since $K(\gamma\pi) = K(\pi) - \log(\gamma)$ for each $\pi \in \mathcal{M}$, $\gamma \geq 0$, it follows that (66) holds for all $\pi \in \mathcal{M}$. □

The following simple result is used in an analysis of the functional K.

Lemma A.5. *For each $\theta \in \mathbb{R}$ the functional $Y(\pi) = M_{\pi,h}(\theta)$ is concave and Gateaux differentiable on $\mathcal{M}$. Its derivative is represented by the function $g_{\pi,h} = \langle \pi, e^{\theta h}\rangle^{-1} e^{\theta h}$, so that*

$$Y(\pi) \leq Y(\pi^0) + \langle \pi - \pi^0, g_{\pi,h}\rangle, \qquad \pi, \pi^0 \in \mathcal{M}.$$

□

The next result from convex analysis is required in the proofs of the major results. We adopt the following notation: $\mathbb{X}$ and $\mathbb{Y}$ denote normed linear spaces; $\mathbb{Y}^*$ denotes the usual dual space of continuous linear functionals on $\mathbb{Y}$; $\mathcal{K}$ denotes a convex subset of $\mathbb{X}$; Ω is a real-valued convex functional defined on $\mathcal{K}$; and the mapping $\Theta\colon \mathbb{X} \to \mathbb{Y}$ is affine.

For a proof of Proposition A.6 see [27, Problem 7, page 236], following [27, Theorem 1, page 224].

Proposition A.6. *Suppose that $\mathbb{Y}$ is finite-dimensional, and that the following two conditions hold:*

(a) *The optimal value κ_0 is finite, where*

$$\kappa_0 := \inf\{\Omega(x)\colon \Theta(x) = 0,\ x \in \mathcal{K}\}.$$

(b) *$0 \in \mathbb{Y}$ is an interior point of the non-empty set $\{y \in \mathbb{Y} : \Theta(x) = y$ for some $x \in \mathcal{K}\}$.*

Then, there exists $y_0^* \in \mathbb{Y}^*$ such that

$$\kappa_0 = \inf\{\Omega(x) + \langle \Theta(x), y_0^*\rangle\colon x \in \mathcal{K}\}$$

□

Below we collect results required in the proof of Theorem 3.2.

Lemma A.7. *Suppose that $\mathcal{H}$ is expressed as (22) for some $r \in (\overline{r}_h, \overline{h})$, and suppose that $\beta := \inf_{\mu\in\mathcal{H}} L(\mu) > 0$. Then, there exists $\mu^* \in \mathcal{H}$, $\pi^* \in \mathbb{P}$, $\theta^* > 0$, and $\lambda^* \in \mathbb{R}^{n+1}$ satisfying,*

(i) $\beta = L(\mu^*) = K(\pi^*) = \inf_{\pi\in\mathbb{P}} K(\pi)$,

(ii) $K(\pi^*) = \sup_{\theta\geq 0}\{\theta r - M_{\pi^*,h}(\theta)\} = \{\theta^* r - M_{\pi^*,h}(\theta^*)\}$,

(iii) $\dfrac{d\mu^*}{d\pi^*} = \dfrac{\exp(\theta^* h)}{\langle \pi, \exp(\theta^* h)\rangle}$,

(iv) $K(\pi^*) = \min_{\pi\in\mathcal{M}} \big\{K(\pi) + \lambda^{*T}\big(\langle \pi, f\rangle - c\big)\big\}$.

Proof. Theorem 3.1 states that $D(\cdot \parallel \cdot)$ is jointly lower semi-continuous. Since $\mathcal{M}_1$ is compact, it follows that an optimizing pair (μ^*, π^*) exists.

To establish (ii) we must show that the supremum over θ in (66) is attained by some $\theta^* \geq 0$. We prove this by contradiction: Suppose that no finite $\theta^* \in \mathbb{R}_+$ exists, so that

$$K(\pi^*) = \lim_{\theta\to\infty} \{\theta r - M_{\pi^*,h}(\theta)\}.$$

From the definition of K in (66) we then obtain the representation,

$$K(\pi^*) = \inf_{\pi\in\mathbb{P}} \Big(\lim_{\theta\to\infty} \{\theta r - M_{\pi,h}(\theta)\}\Big).$$

However, on taking $\pi \in \mathbb{P}$ with support at $\overline{h}$, we have as in the proof of (64) in Lemma A.1,

$$\lim_{\theta\to\infty} \{\theta r - M_{\pi^*,h}(\theta)\} \leq \lim_{\theta\to\infty} \{\theta r - M_{\pi,h}(\theta)\} = \lim_{\theta\to\infty} \{\theta r - (\log(\overline{p}) + \overline{h}\theta)\} = -\infty,$$

where $\overline{p} = \pi\{x : h(x) = \overline{h}\}$. This contradiction shows that there exists $\theta^* < \infty$ as claimed.

Note that we must have $\theta^* > 0$ since $K(\pi^*) = \beta > 0$. Cramér's Theorem A.3 then implies (iii).

It is straightforward to verify that the infimum in (i) meets the conditions of Proposition A.6, from which we obtain (iv). ◻

The proof of the following result is routine calculus.

Lemma A.8. *Let $\pi^0, \pi^1, \mu \in \mathcal{M}$, and define $\pi^\varrho = (1-\varrho)\pi^0 + \varrho\pi^1$ for $\varrho \in [0,1]$. Then,*

$$\frac{d}{d\varrho} D(\mu\|\pi^\varrho)\Big|_{\varrho=0} = 1 - \big\langle \mu, \frac{d\pi^1}{d\pi^0}\mathbb{I}_A\big\rangle,$$

where $A \subset \mathsf{X}$ denotes the support of π^0. ◻

Proof of Theorem 1.4 *Proof of (i):* This is based on Proposition A.6 with the identification,

$$\mathbb{X} = \mathcal{S};\ \mathbb{Y} = \mathbb{R}^{n+1};\ \mathcal{K} = \mathcal{M};\quad \Omega(\pi) = D(\mu \parallel \pi);\ \Theta(\pi) = \langle \pi, f\rangle - c.$$

We now verify the two required assumptions in Proposition A.6: (a) the infimum (16) that defines L must be finite, and (b) the constraint vector c must lie in the interior of the set Δ. The first property is established in Lemma A.1 (i), and the second is guaranteed by Assumption (A1).

Consequently, Proposition A.6 implies the following expression for the worst-case rate function:

$$L(\mu) = \inf_{\pi\in\mathbb{P}} D(\mu \parallel \pi) = \max_{\lambda\in\mathbb{R}^n} \Psi(\lambda), \qquad \text{where} \qquad \Psi(\lambda) := \inf_{\pi\in\mathcal{M}} \Big\{D(\mu \parallel \pi) + \lambda^T(\langle \pi, f\rangle - c)\Big\}.$$

We now obtain an expression for Ψ. Consider $\lambda \in R(f)$ statisfying $\lambda^T f > 0$ a.e. $[\mu]$, and define the positive measure π^λ through $\frac{d\pi^\lambda}{d\mu} = \frac{1}{\lambda^T f}$ Note that this may not be a probability measure, but we will prove that π^λ achieves the minimum in the definition of $\Psi(\lambda)$.

Define $\pi^\varrho := \pi^\lambda + \varrho(\pi - \pi^\lambda)$ for a given $\pi \in \mathcal{M}$, and $\varrho \in [0,1]$. We have,

$$\begin{aligned}\frac{d}{d\varrho}\Big(D(\mu \parallel \pi^\varrho) + \lambda^{\scriptscriptstyle T}(\langle \pi^\varrho, f\rangle - c)\Big)\Big|_{\varrho=0} &= 1 - \big\langle \mu, \frac{d\pi}{d\pi^\lambda}\mathbb{I}_A\big\rangle + \lambda^{\scriptscriptstyle T}\langle \pi - \pi^\lambda, f\rangle \\ &= 1 - \langle \pi, (\lambda^{\scriptscriptstyle T} f)\mathbb{I}_A\rangle + \langle \pi, \lambda^{\scriptscriptstyle T} f\rangle - \langle \mu, 1\rangle \\ &= \langle \pi, (\lambda^{\scriptscriptstyle T} f)\mathbb{I}_{A^c}\rangle.\end{aligned}$$

where the first equality follows from Lemma A.8 with A equal to the support of π^λ. The second equality follows from the definition $(\lambda^{\scriptscriptstyle T} f)\mathbb{I}_A = \frac{d\mu}{d\pi^\lambda}\mathbb{I}_A$. Since $\lambda^{\scriptscriptstyle T} f \geq 0$ this shows that for any $\pi \in \mathcal{M}$,

$$\frac{d}{d\varrho}\Big(D(\mu \parallel \pi^\varrho) + \lambda^{\scriptscriptstyle T}(\langle \pi^\varrho, f\rangle - c)\Big)\Big|_{\varrho=0} \geq 0\,,$$

and hence π^λ achieves the minimum in the expression for $\Psi(\lambda)$. The formula for $\Psi(\lambda)$ when $\lambda \in R(f)$ follows.

To complete the proof of the duality relation we show that $\Psi(\lambda) = -\infty$ in either of the two cases, $\lambda \notin R(f)$ *or* $\lambda^{\scriptscriptstyle T} f \not\geq 0, \mu$-*a.e.*. Indeed, if $\lambda \notin R(f)$ then $\lambda^{\scriptscriptstyle T} f(x_0) < 0$ for some $x_0 \in \mathsf{X}$ satisfying $\mu\{x_0\} = 0$. Let $\pi^\kappa := \mu + \kappa\delta_{x_0}$, where δ_{x_0} is the atom at x_0. Then $D(\mu \parallel \pi^\kappa) = 0$ whereas $\langle \pi^\kappa, \lambda^{\scriptscriptstyle T} f\rangle \downarrow -\infty$ as $\kappa \uparrow \infty$.

In the latter case, in which $\lambda^{\scriptscriptstyle T} f$ is not strictly positive a.e. $[\mu]$, it follows that the set $A := \{x : \lambda^{\scriptscriptstyle T} f(x) \leq 0\}$ has positive μ-measure. Consider the sequence of positive measures π^κ defined through $\frac{d\pi^\kappa}{d\mu} = n\mathbb{I}_A$. Then $D(\mu \parallel \pi^\kappa) = -\log(\kappa)$ and $\langle \pi^\kappa, \lambda^{\scriptscriptstyle T} f\rangle \leq 0$ so that $D(\mu \parallel \pi^\kappa) + \langle \pi^\kappa, \lambda^{\scriptscriptstyle T} f\rangle \downarrow -\infty$ as $\kappa \to \infty$.

Proof of (ii) and (iii): We first show that the infimum in (16) and the supremum in (20) are achieved by a pair $\pi^* \in \mathbb{P}$, $\lambda^* \in R(f)$. The fact that the supremum is achieved follows directly from Proposition A.6, and the existence of an optimizing $\pi^* \in \mathbb{P}$ follows from Theorem 3.1.

Convexity of L follows directly from its formulation as a supremum of linear functionals in part (i). The finiteness of L is proved in Lemma A.1 (i).

To show that $L\colon \mathcal{M}_1 \to \mathbb{R}_+$ is continuous, consider any convergent sequence of probability measures, $\mu^k \xrightarrow{\text{w}} \mu$. From part (i) we know that there exist $\{\lambda^k\} \subset R(f,c)$ and $\{\pi^k\} \subset \mathbb{P}$ such that $\frac{d\mu^k}{d\pi^k} = \lambda^{k\,{\scriptscriptstyle T}} f$, and $L(\mu^k) = D(\mu^k \parallel \pi^k)$ for each k.

Consider any limit point of $\{L(\mu^k)\}$, and a subsequence $\{k_i\}$ such that $\{L(\mu^{k_i})\}$ is convergent to this limit point. Since the sets $\mathbb{P}$ and $R(f,c)$ are compact (the latter from Lemma A.2), we can construct if necessary a further subsequence so that $\lambda^{k_i} \to \lambda$ and $\pi^k \xrightarrow{\text{w}} \pi$, for some $\lambda \in R(f,c)$, and $\pi \in \mathbb{P}$. As in the proof of part (i), it follows that $\frac{d\mu}{d\pi} = \lambda^{\scriptscriptstyle T} f$ and $L(\mu) = D(\mu \parallel \pi)$. Since f is a bounded function, we must have $\lambda^{k_i\,{\scriptscriptstyle T}} f \to \lambda^{\scriptscriptstyle T} f$ uniformly on X.

Consider now the functions $\{(\lambda^{k_i\,{\scriptscriptstyle T}} f)\log \lambda^{k_i\,{\scriptscriptstyle T}} f : i \geq 1\}$. Since $x\log x$ is a continuous function on $\mathbb{R}$, it follows that,

$$(\lambda^{k_i\,{\scriptscriptstyle T}} f)\log \lambda^{k_i\,{\scriptscriptstyle T}} f \to (\lambda^{\scriptscriptstyle T} f)\log \lambda^{\scriptscriptstyle T} f, \qquad \text{uniformly on } \mathsf{X},$$

and, since $\pi^{*k_i} \xrightarrow{\text{w}} \pi^*$, we have

$$\langle \pi^{*k_i}, (\lambda^{k_i\,{\scriptscriptstyle T}} f)\log \lambda^{k_i\,{\scriptscriptstyle T}} f\rangle \to \langle \pi, (\lambda^{\scriptscriptstyle T} f)\log \lambda^{\scriptscriptstyle T} f\rangle.$$

From the identities,

$$L(\mu^{k_i}) = \langle \pi^{*k_i}, (\lambda^{k_i\,{\scriptscriptstyle T}} f)\log(\lambda^{k_i\,{\scriptscriptstyle T}} f)\rangle \qquad \text{and} \qquad L(\mu) = \langle \pi, (\lambda^{\scriptscriptstyle T} f)\log(\lambda^{\scriptscriptstyle T} f)\rangle,$$

we conclude that $L(\mu^{k_i}) \to L(\mu)$. This completes the proof of continuity, and thereby establishes (ii).

To prove part (iii), we begin with the representation,

$$\mathcal{Q}_\beta(\mathbb{P}) = \{\mu \colon L(\mu) < \beta\}.$$

This set is convex and open since the functional L is convex and continuous. We have seen that the infimum in (16) is achieved by some $\pi^* \in \mathbb{P}$, from which it follows that

$$\mathcal{Q}_\beta^+(\mathbb{P}) = \{\mu \colon L(\mu) \le \beta\}.$$

Since L is convex and continuous, $\mathcal{Q}_\beta^+(\mathbb{P})$ is convex and closed. It easily follows from these expressions and continuity of L that $\mathcal{Q}_\beta^+(\mathbb{P})$ is equal to the closure of $\mathcal{Q}_\beta(\mathbb{P})$. ☐

Proof of Theorem 3.2 Part (i) *(Necessity)*: We apply Lemmas A.7 and A.5 to establish the alignment condition (49): Consider any $\pi^0 \in \mathcal{M}$ such that the supremum in (66) is achieved for some $\theta^0 \in \mathbb{R}_+$. We apply the following bound,

$$\beta = K(\pi^*) \ge \theta^0 r - M_{\pi^*,h}(\theta^0) \ge \theta^0 r - M_{\pi^0,h}(\theta^0) - \langle \pi^* - \pi^0, g_0 \rangle = K(\pi^0) - \langle \pi^* - \pi^0, g_0 \rangle,$$

where the first inequality follows from Lemma A.7 (ii), and the second is a consequence of Lemma A.5 with $g_0 := g_{\pi^0,h}$. Consequently, for any $\lambda \in R(f,c)$,

$$\begin{aligned} K(\pi^*) + \lambda^{\scriptscriptstyle T}(\langle \pi^*, f\rangle - c) \ge K(\pi^0) + \lambda^{\scriptscriptstyle T}(\langle \pi^0, f\rangle - c) \\ + \langle \pi^* - \pi^0, \lambda^{\scriptscriptstyle T}(f - c) - g_0 \rangle, \end{aligned}$$

and on combining this with Lemma A.7 (iv) we obtain,

$$\langle \pi^* - \pi^0, \lambda^{*\,{\scriptscriptstyle T}}(f - c) - g_0 \rangle \le 0.$$

Consider the special case $\pi^0 = \pi^* + \epsilon\delta_x$ for $\epsilon > 0$, $x \in \mathsf{X}$, so that the bound above becomes,

$$g_0(x) \le \lambda^{*\,{\scriptscriptstyle T}}(f(x) - c).$$

It is clear that $M_{\pi^0,h} \to M_{\pi^*,h}$ as $\epsilon \downarrow 0$, uniformly for θ in compact subsets of $\mathbb{R}_+$. The function $M_{\pi^*,h}$ is strictly convex on $\mathbb{R}_+$, from which we conclude that $\theta^0 \to \theta^*$ as $\epsilon \downarrow 0$. It then follows that the function $g_0 = g_{\pi^0,h}$ converges to $g_{\pi^*,h} = \langle \pi, e^{\theta^* h}\rangle^{-1} e^{\theta^* h}$ uniformly on X. We conclude that

$$\frac{e^{\theta^* h(x)}}{\langle \pi, e^{\theta^* h}\rangle} \le \lambda^{*\,{\scriptscriptstyle T}}(f(x) - c), \qquad x \in \mathsf{X}.$$

Now consider the special case $\pi^0 = (1-\epsilon)\pi^*$ for $\epsilon > 0$. In this case $\theta^0 = \theta^*$, and we can use identical arguments to conclude that $\langle \pi^*, \lambda^{*\,{\scriptscriptstyle T}}(f - c) - g_*\rangle \le 0$. The representation (49) follows on taking logarithms.

Part (ii) *(Sufficiency)*: From the assumption $\theta^* h \le \theta^* r - \beta + \log(\lambda^{\scriptscriptstyle T} f)$ we obtain from (66),

$$\begin{aligned} K(\pi) &\ge \theta^* r - M_{\pi,h}(\theta^*) \\ &:= \theta^* r - \log(\langle \pi, \exp(\theta^* h)\rangle) \\ &\ge \theta^* r - [\theta^* r - \beta + \log(\langle \pi, \lambda^{\scriptscriptstyle T} f\rangle) \quad = \beta. \end{aligned}$$

Moreover, since we also have $\log \frac{d\mu^*}{d\pi^*} = \beta + \theta^*(h - r)$ and $\langle \mu^*, h \rangle = r$, we can conclude that this lower bound is achieved when $\pi = \pi^*$:

$$K(\pi^*) \leq D(\mu^* \| \pi^*) = \langle \mu^*, \beta + \theta^*(h - r) \rangle = \beta.$$

It follows that for any $\mu \in \mathcal{H}$, $\pi \in \mathbb{P}$,

$$D(\mu \| \pi) \geq K(\pi) \geq \beta = K(\pi^*) = L(\mu^*).$$

Consequently, $\mathcal{Q}_\beta(\mathbb{P}) \subset \mathcal{H}^1$ and $\mathcal{Q}_\beta^+(\mathbb{P}) \subset \mathcal{H} \cup \mathcal{H}^1$. Since $\mu^* \in \mathcal{H} \cap \mathcal{Q}_\beta^+(\mathbb{P})$, the hyperplane $\mathcal{H}$ must be a supporting hyperplane for $\mathcal{Q}_\beta^+(\mathbb{P})$. □

The two lemmas below are consequences of Theorem 3.2. The first provides a characterization of an extremal distribution in terms of the threshold function h, the constraints $\{f_i, c_i\}$, and the value $r \in \mathbb{R}$.

Lemma A.9. *A necessary and sufficient condition for a distribution $\pi^* \in \mathbb{P}$ to be $(h, r, +)$-extremal for some $r \in (\overline{r}_h, \overline{h})$ is that there exist $\pi^* \in \mathcal{H}^0 \cap \mathbb{P}$, $\mu^* \in \mathcal{H}$, $\lambda \in R(f)$, and $\theta^*, b_0 > 0$ such that*

$$\exp(\theta^* h) \begin{cases} = b_0 \, \lambda^T f = b_0 \frac{d\mu^*}{d\pi^*}, & \text{a.e. } [\pi^*] \\ \leq b_0 \, \lambda^T f & \text{everywhere.} \end{cases} \tag{67}$$

Proof. The fact that $\mathcal{H}$ is a supporting hyperplane for $\mathcal{Q}_{\beta^*}^+(\mathbb{P})$, together with Theorem 3.2, implies that, for some $\lambda \in R(f)$ and $\theta^* > 0$, we must have

$$h - r \begin{cases} = \frac{1}{\theta^*}(\log \lambda^T f - \beta^*) = \frac{1}{\theta^*}(\log \frac{d\mu^*}{d\pi^*} - \beta^*), & \text{a.e. } [\pi^*] \\ \leq \frac{1}{\theta^*}(\log \lambda^T f - \beta^*) & \text{everywhere.} \end{cases}$$

We conclude that the relation (67) is a necessary condition for π^* to be an extremal distribution, where $b_0 := \exp(\theta^* r - \beta^*)$.

Conversely, if π^*, μ^* satisfy (67) along with $D(\mu^* \parallel \pi^*) = \beta^*$, then from Theorem 3.2, the hyperplane $\mathcal{H}$ supports $\mathcal{Q}_{\beta^*}^+(\mathbb{P})$, and therefore the pair $\{\pi^*, \mu^*\}$ solves (26). Thus (67) is also a sufficient condition for π^* to be an extremal distribution. □

Lemma A.10. *Let $\mathbb{P}$ be a moment class that satisfies Assumption (A1). Suppose that $\pi^* \in \mathbb{P}$ is $(h, r, +)$-extremal for some $r \in (\overline{r}_h, \overline{h})$, and let $\theta^* > 0$ be the constant given in Lemma A.9. Then, π^* is also an optimizer of the infinite-dimensional linear program (7) that defines the worst-case moment-generating function, with $\theta = \theta^*$.*

Proof. From (67) it follows that for any $\pi \in \mathbb{P}$ we have

$$\langle \pi, \exp(\theta^* h) \rangle \leq \langle \pi, b_0 \lambda^T f \rangle = b_0 \lambda^T c,$$

with equality when $\pi = \pi^*$. Thus π^* solves (7). □

Proof of Theorem 1.5 The proof follows from Lemma A.10 and Lemma A.7 (ii): From these results and maximality of $\overline{M}_h$ we have,

$$\underline{I}_h(r) \geq \theta r - M_{\pi^*, h}(\theta) \geq \theta r - \overline{M}_h(\theta), \qquad \theta \geq 0,$$

and all inequalities become equalities when $\theta = \theta^*$. □

References

[1] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57:33–45, 1962.

[2] D.P. Bertsekas. *Convex Analysis and Optimization*. Atena Scientific, Cambridge, Mass, 2003. with Angelia Nedic and Asuman E. Ozdaglar.

[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] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 1st edition, 2004.

[5] F. Brichet and A. Simonian. Conservative Gaussian models applied to measurement-based admission control. In *Proceedings of IWQoS*, pages 68–71, Napa, CA, May 1998.

[6] M. Chiang and S. Boyd. Geometric programming duals of channel capacity and rate distortion. *IEEE Transactions on Information Theory*, 50(2):245–258, 2004.

[7] A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications*. Springer-Verlag, New York, second edition, 1998.

[8] 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.

[9] A. Ganesh, N. O'Connell, and D. Wischik. *Big queues*, volume 1838 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.

[10] R. Gibbens and F. Kelly. Measurement-based connection admission control, 1997. In International Teletraffic Congress 15, Jun. 1997.

[11] G. C. Goodwin and K. S. Sin. *Adaptive Filtering Prediction and Control*. Prentice Hall, Englewood Cliffs, NJ, 1984.

[12] S. G. Henderson, S. P. Meyn, and V. B. Tadić. Performance evaluation and policy selection in multiclass networks. *Discrete Event Dynamic Systems: Theory and Applications*, 13(1-2):149–189, 2003. Special issue on learning, optimization and decision making (invited).

[13] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[14] W. Hoeffding. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, 36:369–408, 1965.

[15] J. Huang, S. Meyn, and M. Medard. Error exponents for channel coding and signal constellation design. *Journal of Selected Areas in Comm. Special Issue on Nonlinear Optimization of Communication Systems (submitted)*, 2005.

[16] J. Huang and S. P. Meyn. Characterization and computation of optimal distribution for channel coding. *IEEE Trans. Inform. Theory*, 51(7):1–16, 2005.

[17] J. Huang, C. Pandit, S. Meyn, M. Medard, and V. Veeravalli. Entropy, inference, and channel coding. In Prathima Agrawal, Matthew Andrews, Philip J. Fleming, George Yin, and Lisa Zhang, editors, *Proceedings of the Summer Workshop on Wireless Networks (To appear.)*, IMA volumes in Mathematics and its Applications, New York, 2005. Springer-Verlag.

[18] P. J. Huber and V. Strassen. Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.*, 1:251–263, 1973.

[19] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: nonlinear programming approaches. *Comm. Statist. Stochastic Models*, 6(2):259–281, 1990.

[20] M. A. Johnson and M. R. Taaffe. An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queueing Systems Theory Appl.*, 8(2):129–147, 1991.

[21] 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.

[22] J.M.B. Kemperman. The general moment problem, a geometric approach. *Annals of Mathematical Statistics*, 39:93–122, 1968.

[23] I. Kontoyiannis, L.A. Lastras-Montaño, and S. P. Meyn. Relative entropy and exponential deviation bounds for general Markov chains. In *Proceedings of the IEEE International Symposium on Information Theory, Adelaide, Australia, 4-9 September.*, 2005.

[24] I. Kontoyiannis and S. P. Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.*, 13:304–362, 2003. Presented at the INFORMS Applied Probability Conference, NYC, July, 2001.

[25] I. Kontoyiannis and S. P. Meyn. Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electron. J. Probab.*, 10(3):61–123 (electronic), 2005.

[26] 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.

[27] D. G. Luenberger. *Optimization by vector space methods*. John Wiley, 1969.

[28] D.G. Luenberger. *Linear and nonlinear programming*. Kluwer Academic Publishers, Norwell, MA, second edition, 2003.

[29] 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.

[30] S. P. Meyn. Dynamic safety-stocks for asymptotic optimality in stochastic networks. *Queueing Syst. Theory Appl.*, 50:255–297, 2005.

[31] 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.

[32] C. Pandit and S. P. Meyn. Robust measurement-based admission control using Markov's theory of canonical distributions. Submitted to the IEEE Trans. on Information Theory. Conference version presented at the Conference on Information Sciences and Systems, 2003, 2003.

[33] C. Pandit, S. P. Meyn, and V. V. Veeravalli. Asymptotic robust Neyman-Pearson hypothesis testing based on moment classes. In preparation, 2005.

[34] 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.

[35] 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.

[36] 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.

[37] I. N. Sanov. On the probability of large deviations of random magnitudes. *Mat. Sb. N. S.*, 42 (84):11–44, 1957.

[38] 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. Papers from the AMS-IMS-SIAM Joint Summer Research Conference held in Seattle, Washington, July 1991.

[39] A. Shapiro and A. Nemirovski. Duality of linear conic problems. *Published electronically in:* Optimization Online, pages 1–27, 2003.

[40] Adam Shwartz and Alan Weiss. *Large deviations for performance analysis.* Stochastic Modeling Series. Chapman & Hall, London, 1995. Queues, communications, and computing, With an appendix by Robert J. Vanderbei.

[41] J. E. Smith. Generalized Chebychev inequalities: theory and applications in decision analysis. *Operations Res.*, 43(5):807–825, 1995.

[42] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyshev bounds via semidefinite programming. Submitted to SIAM Review, Problems and Techniques Section, January 2004.

[43] V. V. Veeravalli, T. Başar, and H. V. Poor. Minimax robust decentralized detection. *IEEE Transactions on Information Theory*, 40(1):35–40, 1994.

[44] W. Whitt. Bivariate distributions with given marginals. *Ann. Statist.*, 4(6):1280–1289, 1976.

[45] Ofer Zeitouni and Michael Gutman. On universal hypotheses testing via large deviations. *IEEE Trans. Inform. Theory*, 37(2):285–290, 1991.