Extremal Distributions
and
Worst-Case Large-Deviation Bounds∗

Charuhas Pandit† Sean Meyn‡
September 26, 2004

Abstract

At the end of the nineteenth century, A. A. Markov developed foundations for worst-case Chebyshev inequalities based on polynomial moment constraints on the marginal distribution for an i.i.d. process on a closed, bounded interval in \( \mathbb{R} \). This paper contains the following extensions of these ideas: An i.i.d. process \( X \) is considered on a compact metric space \( X \). Its marginal distribution \( \pi \) is assumed to lie in a moment class of the form,

\[
\mathcal{P} = \{ \pi : \langle \pi, f_i \rangle = c_i, \quad i = 1, \ldots, n \},
\]

where \( \{f_i\} \) are real-valued, continuous functions on \( X \), and \( \{c_i\} \) are constants.

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

\[
L(\mu) := \inf_{\pi \in \mathcal{P}} D(\mu \parallel \pi) = \sup_{\lambda \in R(f,c)} \langle \mu, \log(\lambda^T f) \rangle,
\]

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

Properties of extremal distributions are also developed. These are interpreted as boundary points for the sublevel sets of \( L \). Extremal distributions are a generalization of Markov’s canonical distributions, and it is shown that they provide a solution to a worst-case one-dimensional large-deviations problem.
1 Introduction & Background

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

$$P := \{\pi \in M_1 : \langle \pi, f_i \rangle = c_i, \quad i = 1, \ldots, n\}, \quad (1)$$

where $M_1$ 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 $\pi$.

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 [1]; admission control [2, 3, 4]; queueing theory [5, 6, 7]; and other applications.

For a moment class of this form, and a given function $g \in C(X)$, the map $\pi \mapsto \langle \pi, g \rangle$ defines a continuous linear functional on $M_1$. Consequently, the following maximization may be viewed as a linear program,

$$\max_{\pi \in P} \langle \pi, g \rangle \quad (2)$$

The value of this (infinite-dimensional) linear program provides a bound on the mean of $g$ that is uniform over $\pi \in P$. 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. Kreǐn in 1959 describes many of Markov’s original results [8]. Since then, these ideas have been developed in various directions [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19].

The present paper concerns large-deviations bounds that are uniform across a moment class. For a given function $h \in C(X)$, and any $r \geq \langle \pi, h \rangle$, the simplest large-deviations bound is Chernoff’s bound:

$$P\{S_N \geq r\} \leq \exp(-NI_{\pi,h}(r)), \quad N \geq 1, \quad (3)$$

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 $\pi$. Denoting the log moment-generating function as,

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

the rate-function is equal to the convex dual,

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

The infimum in (5) may be restricted to $\theta \geq 0$ since $r \geq \langle \pi, h \rangle$.

Consider now the problem of obtaining bounds on the exceedance probability in (3) when the distribution $\pi$ is not known exactly, but is known to lie in the moment class $P$. In this case we are interested in obtaining an upper bound on the large deviations probability that is uniform across all distributions $\pi \in P$.

To motivate the results obtained in this paper we present some known results in the special case of polynomial constraint functions.
1.1 Markov’s canonical distributions

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

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

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

$$m_{h}(\theta) := \max_{\pi \in \mathbb{P}} \langle \pi, \exp(\theta h) \rangle, \quad \theta \in \mathbb{R}. \tag{7}$$

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 [8] that there is a single probability distribution $\pi^* \in \mathbb{P}$ that optimizes (7) simultaneously for every $\theta \in \mathbb{R}_+$. The probability distribution $\pi^*$ 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 $\pi^*$ is a discrete distribution with exactly $\left\lceil \frac{n}{2} \right\rceil + 1$ points of support. Moreover, if $n$ is even, then the end-point 1 lies in the support of $\pi^*$. If $n$ is odd, then the end-points $\{0, 1\}$ each lie in the support of $\pi^*$.

\[\square\]

It can be shown that finding the distribution $\pi^*$ is equivalent to solving an $n^{th}$ degree polynomial. Consequently, analytical formulae for $\pi^*$ 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 $\pi^*$ 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 [10], and the case $n = 2$ was considered by Bennett [9] 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,

\[ I(r) := \inf \{ I_\pi(r) : \pi \in \mathcal{P} \}, \quad r \in \mathbb{R}, \tag{10} \]

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

\[ I_{\pi^*}(r) = I(r), \quad r \geq c_1. \tag{11} \]

Consequently, the universal Chernoff bound holds,

\[ \mathbb{P}\{S_N \geq r\} \leq \exp(-NI_{\pi^*}(r)), \quad \pi \in \mathcal{P}, \ N \geq 1, \ r \geq c_1. \]

\[ \square \]

1.2 Main results

How can Theorem 1.1 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 definitions and results from [20]. For two distributions \( \mu, \pi \in \mathcal{M}_1 \), the relative entropy, or Kullback-Leibler divergence is defined as,

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

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

\[ N_g(s, \epsilon) := \{ \mu \in \mathcal{S} : |\mu(g) - s| < \epsilon \}, \quad \text{for real-valued } g \in C(X), \ s \in \mathbb{R}, \ \epsilon > 0 \}. \tag{12} \]

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

The sequence of empirical distributions is defined by,

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

We then have the well-known limit theorem,

Theorem 1.3. (Sanov’s Theorem for Empirical Measures) Suppose that \( X \) is i.i.d. with marginal distribution \( \pi \) 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

\[ I(\mu) := D(\mu \parallel \pi), \quad \mu \in \mathcal{M}_1. \]  

(14)

Consequently, for any \(E \in \mathcal{F}\),

\[- \inf_{\mu \in E^o} I(\mu) \leq \lim\inf_{N \to \infty} N^{-1} \log L_N(E) \leq \lim\sup_{N \to \infty} N^{-1} \log L_N(E) \leq - \inf_{\mu \in \overline{E}} I(\mu),\]

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

\[\square\]

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,

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

(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 explicit expression for the worst-case rate-function \(L : \mathcal{M}_1 \to \mathbb{R}\) defined as,

\[ L(\mu) := \inf_{\pi \in \mathcal{P}} D(\mu \parallel \pi). \]  

(16)

Before presenting this result, we fix some notation. Let \(\{f_1, \ldots, f_n\}\) be continuous functions on \(X\), and \(\{c_1, \ldots, c_n\}\) the constants in the definition (1). We let \(f : X \to \mathbb{R}^{n+1}\) denote the vector of functions \((1, f_1, \ldots, f_n)^T\), and write \(c := (1, c_1, \ldots, c_n)^T \in \mathbb{R}^{n+1}\).

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

\[ Q_\beta(\pi) := \{\mu \in \mathcal{M}_1 : D(\mu \parallel \pi) < \beta\}, \]

\[ Q_\beta^+(\pi) := \{\mu \in \mathcal{M}_1 : D(\mu \parallel \pi) \leq \beta\}. \]  

(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 \(\mathcal{P}\):

\[ Q_\beta(\mathcal{P}) = \bigcup_{\pi \in \mathcal{P}} Q_\beta(\pi) \quad \text{and} \quad Q_\beta^+(\mathcal{P}) = \bigcup_{\pi \in \mathcal{P}} Q_\beta^+(\pi). \]  

(18)

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

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

(19)
A version of Theorem 1.4 appears as Proposition 2.2.1 in the dissertation [21]. A proof is included in Section 2.3.

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

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

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

where

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

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

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

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

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

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

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

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

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 $h \in C(X)$. 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 \( L_h \) 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 \([\overline{r}_h, \overline{h}] \), where

\[
\underline{r}_h = \sup \{ r : \mathcal{H}(r) \cap \mathbb{P} \neq \emptyset \}, \quad \overline{r}_h = \max \{ h(x) : x \in X \}; \\
\underline{h} = \min \{ h(x) : x \in X \}.
\]

(24)

An interpretation of these constants is illustrated in Figure 2, and in Proposition 2.3 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 \([\underline{r}_h, \overline{r}_h] \):

\[
L_h(r) := \inf_{\pi \in \mathbb{P}} \inf_{\mu \in \mathcal{H}_1} D(\mu \parallel \pi), \quad r \in [\underline{r}_h, \underline{h}].
\]

(25)

This gives rise to the notion of extremal distributions:

Given a moment class \( \mathbb{P} \), a function \( h \in C(X) \), and \( r \in (\underline{r}_h, \overline{r}_h) \), 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 \( r \in (\underline{r}_h, \overline{r}_h) \) ensures that \( L_h(r) > 0 \). A \((h, r, -)-extremal\) distribution is defined analogously for \( r \in (\overline{h}, \underline{r}_h) \).

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

The paper [17] uses the exact same terminology for distributions that solve a particular infinite dimensional linear program. Although the setting is very different, Theorem 2.2 shows that the definition used here is consistent with the definition of extremal distributions introduced in [17].

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

\[
L_h(r) = \inf_{\pi \in \mathbb{P}} \inf_{\mu \in \mathcal{H}_1} D(\mu \parallel \pi), \quad r \in (\underline{r}_h, \overline{r}_h).
\]

(26)
Which is equivalently expressed,

$$L_h(r) = \sup\{\beta : Q^+_{\beta}(P) \cap H = \emptyset\}, \quad r \in (r_h, h).$$  \hspace{1cm} (27)

This follows from the geometry illustrated in Figure 1.

The set $H$ forms a supporting hyperplane for $Q^+_{\beta}(P)$, passing through distributions $\mu^*$ in the intersection $Q^+_{\beta}(P) \cap H$. Theorem 1.4 asserts that there exists $\pi^* \in P$ such that $D(\mu^* \parallel \pi^*) = \beta^*$. The pair of probability distributions $\{\mu^*, \pi^*\}$ solve (26), and $\pi^*$ is an extremal distribution.

The remainder of the paper is organized as follows: Section 2.1 contains a development of the geometry illustrated in Figure 1. Based on this structure, we characterize extremal distributions in terms of the constraint functions $\{f_i\}$ and the threshold function $h$. Section 2.1 also contains some illustrative numerical examples. Extremal distributions are investigated in Section 2.2. Here it is shown that they can be assumed discrete without loss of generality. Section 2.3 contains proofs of the major results.

Conclusions are contained in Section 3, along with a description of possible future directions and potential applications for this research.

2 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 $M_1$ denotes the space of (Borel) probability measures on on the Borel $\sigma$-algebra $\mathcal{B}$, endowed with the weak-topology.

We begin with an examination of the functional $L$, and the associated divergence sets defined in (18).

2.1 Convex geometry of divergence sets

A characterization of supporting hyperplanes is provided in the next result. For $h \in C(X)$ and $\pi \in M_1$, we denote by $\overline{h}_{\pi}$ the essential supremum of $h$ under $\pi$.

**Theorem 2.1.** (Identification of Supporting Hyperplanes) Suppose that $\mu^* \in \partial Q^+_{\beta}(P)$, and that $H$ is a supporting hyperplane for the divergence set $Q^+_{\beta}(P)$ at $\mu^*$ in the sense that $\mu^* \in Q^+_{\beta}(P) \cap H$, and $Q^+_{\beta}(P) \subset H^0$.

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

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

$$h - r \begin{cases} = \frac{1}{\theta^*}(\log \lambda^T f - \beta) = \frac{1}{\theta^*}(\log \frac{d\mu^*}{d\pi^*} - \beta), & \text{a.e. } [\pi^*] \\ \leq \frac{1}{\theta^*}(\log \lambda^T f - \beta) & \text{everywhere.} \end{cases}$$  \hspace{1cm} (28)

(ii) Conversely, if $\{\mu^*, \pi^*, \theta^*, \lambda\}$ satisfy (28) along with $D(\mu^* \parallel \pi^*) = \beta$, then $H$ supports $Q^+_{\beta}(P)$, with $Q_{\beta}(P) \subset H^0$. 


There are several useful corollaries to this result, especially when considering extremal distributions in the next subsection. The following result shows that given any supporting hyperplane for $Q_\beta(P)$ passing through $\mu \in \partial Q_\beta(P)$, one can construct a supporting hyperplane $\hat{H}$ that has a special form, and also passes through $\mu$.

**Corollary 2.1.** Let $P$ be a moment class that satisfies Assumption (A1). For a given $r \in (\bar{r}_h, \bar{r}_h)$ define $\beta = I_h(r)$, so that the set $H$ defined in (22) is a supporting hyperplane for $Q_\beta(P)$, with $Q_\beta(P) \subset H^1$. Then $\hat{h} = \log(\lambda^T f)$ is a continuous function on $X$, where $\lambda \in \mathbb{R}(f)$ is given in Theorem 2.1. Moreover, the set

$$
\hat{H} := \{ \mu \in M_1 : \langle \mu, \hat{h} \rangle = \beta \}
$$

is also a supporting hyperplane for $Q_\beta(P)$, with

$$
Q_\beta(P) \subset \hat{H}^0 := \{ \mu \in M_1 : \langle \mu, \hat{h} \rangle < \beta \};
$$

$$
H^1 \subset \hat{H}^1 := \{ \mu \in M_1 : \langle \mu, \hat{h} \rangle > \beta \}.
$$

**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 $x \in X$. Since $H$ supports $Q_\beta(P)$, it must satisfy the necessary conditions of Theorem 2.1 for some $\theta^* > 0$ and $\lambda \in R(f)$. From Theorem 2.1, we know that $\hat{h}$ is bounded below by $h$, which is bounded. Therefore we do have $\lambda^T f > 0$ on $X$.

With $\theta^* := 1$, it is clear that $\{\hat{h}, \mu^*, \pi^*, \hat{\theta}^*, \lambda\}$ satisfy the sufficient conditions of Theorem 2.1. Thus the hyperplane $\hat{H}$ forms a supporting hyperplane for $Q_\beta(P)$ with $Q_\beta(P) \subset \{ \mu \in M_1 : \langle \mu, \hat{h} \rangle < \beta \}$. Moreover, since $\theta^*(h - r) \leq \hat{h} - \beta$ everywhere, we have

$$
H_1 = \{ \mu \in M_1 : \langle \mu, h - r \rangle > 0 \} \subset \{ \mu \in M_1 : \langle \mu, \hat{h} - \beta \rangle > 0 \} = \hat{H}_1
$$

□

We now provide illustrations of the alignment conditions (28) through numerical examples. In each example below the state space is taken to be the finite set, $X = \{0.01, 0.02, \ldots, 0.99, 1\}$.

From Theorem 1.4 it follows that $L$ can be expressed as the maximum,

$$
L(\mu) = \max_{\lambda \in R(f,c)} \langle \mu, \log \lambda^T f \rangle, \quad \mu \in M_1,
$$

where

$$
R(f,c) := R(f) \cap \{ \lambda : \lambda^T c = 1 \}. \quad (29)
$$

This is a concave program that can be solved in various different ways. In the experiments illustrated below a logarithmic penalty approach was adopted, wherein the above optimization was replaced by

$$
sup\{ \langle \mu, \log \lambda^T f \rangle + \epsilon(\nu, \log \lambda^T f) : \lambda \in \mathbb{R}^{n+1}, \lambda^T c = 1 \} \quad (30)
$$

9
where $\nu$ denotes the uniform distribution on $X$, and $\epsilon \ll 0.01$. The above maximization can be solved using the standard Newton-Raphson method since the state space is not large. For more complex problems with a larger state space, it may be preferable to apply the conjugate gradient algorithm (see [22, 23] for definitions and alternative approaches.) See also [17], where algorithms for the multi-dimensional polynomial moment problem are constructed, and [24, 25], where convex-analytic techniques are developed for computation of mutual information.

**Case I: $\mu$ binary**  Consider the symmetric binary distribution supported on $\{0.01, 0.99\}$, define $f(x) = (1, x, \ldots, x^n)^T$, for each $n \geq 1$, $x \in X$, and $c = \langle \nu, f \rangle$ with $\nu$ the uniform distribution on $X$. That is, for each $i$ we define $c_i = \langle \nu, x^i \rangle = 100^{-1}\sum_{k=1}^{100}(k/100)^i$, $i \geq 1$.

Shown in Figure 3 are results from numerical calculation of $L$ for $n = 1, 3, 10$ and 20. When $n = 1$ we have $c = (1, 0.5)^T$, so that $\mu \in \mathcal{P}$, and hence $L(\mu) = 0$. For each $n \geq 2$ we have $\mu \notin \mathcal{P}$ and, as shown in the figure for three values of $n$, the worst-case divergence $L(\mu)$ is strictly positive. Also shown in Figure 3 is the $n$th order polynomial $\lambda^T f$ in each case, and the roots of this polynomial contained in $X$. From Theorem 1.4 we know that $L(\mu) = D(\mu \parallel \pi^*)$ for some $\pi^* \in \mathcal{P}$ with $\frac{d\mu}{d\pi^*} = \lambda^T f$. It follows that $\pi^*$ is supported on the union,

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

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

Consider now the case where the functions $\{f_i\}$ are defined via the elements of $X$: $f_i(x) = \mathbb{1}_{x_i}$, $i = 1, \ldots, n$, where $x_i = i/100$ for $i = 1, \ldots, 100$. In this case, for each $n < 100$ the distribution $\pi^*$ is given by,

$$\pi^*(x_i) = \frac{1}{100}, \quad 1 \leq i \leq n; \quad \pi^*(0.99) = \frac{100 - n}{100}.$$
and consequently we have

\[
L(\mu) = \mu(0.1) \log \left( \frac{\mu(0.1)}{\pi^*(0.1)} \right) + \mu(0.99) \log \left( \frac{\mu(0.99)}{\pi^*(0.99)} \right)
= \frac{1}{2} \left( \log(50) + \log(50(100 - n)^{-1}) \right).
\]

Figure 4 shows a comparison of \( L(\mu) \) using these functions, and using polynomials as in Figure 3. When the number of moment constraints \( n \) is large, one would expect \( L(\mu) \) to be close to the relative entropy \( D(\mu \parallel \nu) \), since the moment class \( \mathbb{P} \) corresponds to a neighborhood of \( \nu \). From Figure 3 we see that, for each \( n \geq 1 \), the moment class defined using indicator functions performs strictly better than the moment class defined using polynomials, in the sense that the corresponding value of \( L \) is closer to \( D(\mu \parallel \nu) \).

Case II: \( \mu \) uniform
Consider now the case where \( \mu \) is uniform on the first fifty points, \( \{0.01, 0.02, \ldots, 0.50\} \), and define for each \( n \geq 1 \) the moment vector \( c \in \mathbb{R}^{n+1} \) as above.

When the constraint functions \( f_i \) are defined as indicator functions as above, \( f_i(x) = \mathbb{I}_{x_i} \), \( i = 1, \ldots, n \), then for \( n < 50 \) the optimizer \( \pi^* \) in (20) is given by,

\[
\pi^*(x_i) = \frac{1}{100}, \ 1 \leq i \leq n; \quad \pi^*(x_i) = \frac{100 - n}{100}, \ n < i \leq 100.
\]

Consequently, in this case we obtain through routine calculations,

\[
L(\mu) = \log(2) - \left( \frac{50 - n}{50} \right) \log \left( \frac{100 - n}{50 - n} \right).
\]

Figure 5 shows a comparison of \( L(\mu) \) as a function of \( n \) using quadratics and indicator functions. For \( n \leq 48 \) it is apparent that the moment class defined using indicator functions performs strictly worse.

These results illustrate the following issue in approximation: Suppose that we wish to approximate a distribution \( \nu \) using a moment class \( \mathbb{P} \) so that \( L(\mu) \approx D(\mu \parallel \nu) \) for \( \mu \in \mathcal{M}_1 \). In these examples it is seen that the choice of functions \( \{f_i\} \) defining \( \mathbb{P} \) that provide an accurate approximation depends upon the distribution \( \mu \).
Figure 4: Comparison of $L(\mu)$ for the moment class used in the experiment illustrated in Figure 3 for $n = 1, \ldots, 50$, and using the moment class defined using indicator functions. Here $\mu$ is the symmetric binary distribution supported on \{0.01, 0.99\}.

Figure 5: Comparison of $L(\mu)$ for the moment class used in the experiment illustrated in Figure 3 for $n = 1, \ldots, 50$, and using the moment class defined using indicator functions. Here $\mu$ is the uniform distribution on \{0.01, 0.02, \ldots, 0.50\}.
2.2 Extremal distributions

We have seen that $L_h$ can be computed in (25) by first constructing the worst-case rate-function $\mathcal{M}_1 \rightarrow \mathbb{R}_+$, and then applying the contraction principle. Although the theory is less elegant, it is worthwhile to consider the worst-case log moment-generating function $M_h := \log(\mathcal{M}_h)$, where $\mathcal{M}_h$ is defined in the linear program (7). We show in this section that the convex dual of $M_h$ is precisely $L_h$.

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,

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

where $\lambda^T f \geq g$ means $\lambda^T f(x) \geq g(x)$ for all $x \in X$. 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 (31). See [26, 27] and Theorem 2.11 below.

In this section we focus on the specific linear program that defines $m_h$ in (7). We first present the following saddle-point property for the worst-case one-dimensional rate-function.

**Theorem 2.2.** (Saddle-Point Property) For any $r \in (\overline{r}_h, \overline{h}_h)$, there exists $\pi^* \in \mathcal{P}$ and $\theta^* < \infty$ such that

$$
L_h(r) = I_{\pi^*, h}(r) = [\theta^* r - M_{\pi^*, h}(\theta^*)] = \min_{\pi \in \mathcal{P}} \max_{\theta \geq 0} [\theta r - M_{\pi, h}(\theta)]
$$

$$
= \max_{\theta \geq 0} \min_{\pi \in \mathcal{P}} [\theta r - M_{\pi, h}(\theta)] = [\theta^* r - M_h(\theta^*)].
$$

The theorem provides the following alternate expression for the worst-case rate-function defined in (25):

$$
L_h(r) = \max_{\theta \geq 0} [\theta r - M_h(\theta)], \quad r \in (\overline{r}_h, \overline{h}_h).
$$

Recall that in the setting of Theorem 1.1 we have

$$
L_h(r) = \max_{\theta \geq 0} [\theta r - M_{\pi^*, h}(\theta)], \quad r > c_1,
$$

with $\pi^*$ independent of $r$. When the assumptions of Theorem 1.1 are relaxed then this uniform optimality no longer holds. Shown in Figure 6 are numerical results obtained using $h(x) = 2(x - 1) - x \sin(2\pi x)$, $x \in [0, 1]$. The moment class $\mathcal{P}$ 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 $\pi^*$ that maximizes (7) was computed, and the figure shows the log moment-generating function $M_{\pi^*, h}$. As required by Theorem 2.2, the functions $M_h$ and $M_{\pi^*, h}$ coincide at $\theta = \theta^*$. However, the inequality $M_h(\theta) \geq M_{\pi^*, h}(\theta)$ is strict for $\theta \in (0, \theta^*)$ and $\theta \in (\theta^*, 5]$.

Theorem 2.2 provides justification for the correspondences illustrated in Figure 2.
The worst-case log moment-generating function $\overline{M}_h$

**Proposition 2.3.** The worst-case log moment-generating function satisfies,

(i) $\frac{d^+}{d\theta} \overline{M}_h (0) = \overline{r}_h$, 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) $\lim_{\theta \to \infty} \frac{d^+}{d\theta} \overline{M}_h (\theta) = \overline{I}.$

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 $\pi^*$ is a discrete measure. The following theorem establishes that without loss of generality, an extremal distribution can be assumed discrete.

**Theorem 2.4.** (Discrete extremal distributions) Under Assumption (A1) suppose that $h$ is continuous, and that $r \in (\overline{r}_h, \overline{I})$. Then there exists a probability distribution $\pi^o \in \mathbb{P}$ that is discrete, with no more than $n + 2$ points of support, and is also $(r, h, +)$-extremal.

**Proof.** Theorem 2.2 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 $I_{\pi^*, h}(r) = I(r)$. 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 $\pi^o \in \mathbb{P}$ that satisfies $M_{\pi^o, h}(\theta^*) = M_{\pi^*, h}(\theta^*) = \overline{M}_h(\theta^*)$, and also the consistent derivative constraint,

$$\frac{d}{d\theta} M_{\pi^o, h}(\theta^*) = r. \quad (33)$$
It will then follow that \( I_{\pi^\circ, h}(r) = I_{\pi, h}(r) = \overline{I}(r) \), which is the desired conclusion.

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

\[
\frac{d}{d \theta} M_{\pi, h}(\theta) = \frac{\langle \pi, h e^{\theta h} \rangle}{\langle \pi, e^{\theta h} \rangle}, \quad \theta \in \mathbb{R}.
\]

Consequently, equation (33) is expressed as the equality constraint \( \langle \pi, h e^{\theta^* h} \rangle = r \langle \pi, e^{\theta^* h} \rangle \).

Consider then the linear program,

\[
\begin{align*}
\text{max} & \quad \langle \pi, \exp(\theta^* h) \rangle \\
\text{st.} & \quad \langle \pi, f_i \rangle = c_i, \quad i = 0, \ldots, n \\
& \quad \langle \pi, h e^{\theta^* h} - r e^{\theta^* h} \rangle = 0.
\end{align*}
\]

This is feasible since the \((h, r, +)-extremal distribution \pi^* 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 constraints, an extreme point may have no more than \( n + 2 \) points of support. 

### 2.3 Proofs of the main results

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 \( \overline{L}_h \).

**Lemma 2.5.** Let \( \mathbb{P} \) be a moment class that satisfies Assumption (A1). Then,

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

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

(ii) The function \( I_h \) is uniformly bounded on \([h, \overline{h}]\):

\[
\sup_{r \in [h, \overline{h}]} I_h(r) < \infty.
\]

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

Under (A1) the vector \( c \) lies in the interior of \( \Delta \). This assumption and the uniform bound on \( c_\mu \) implies that there exists \( \epsilon > 0 \) such that,

\[
\frac{c - \epsilon c_\mu}{1 - \epsilon} \in \Delta, \quad \mu \in \mathcal{M}_1.
\]

By the definition of \( \Delta \), 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

\[
D(\mu \parallel \pi^\epsilon) \leq - \log \epsilon < \infty
\]

Thus \( L(\mu) = \inf_{\pi \in \mathbb{P}} D(\mu \parallel \pi) \leq |\log \epsilon| \) for all \( \mu \in \mathcal{M}_1 \), and this establishes (i).
Since $h$ is continuous we can find $\bar{x}, \underline{x} \in X$ such that $h(\bar{x}) = \overline{h}$ and $h(\underline{x}) = \underline{h}$. Moreover, exactly as in the construction of $\pi^*$ above, we can construct $\pi \in P$ such that,

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

We then have,

$$M_{\pi, h}(\theta) \geq \log(p e^{\theta \underline{h}} + \overline{p} e^{\theta \overline{h}}), \quad \theta \in \mathbb{R}.$$  

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

$$\sup_{\theta \geq 0} \theta r - M_{\pi, h}(\theta) \leq \sup_{\theta \geq 0} \left[ \theta (r - \overline{h}) - \log(\overline{p}) \right] \leq |\log(\overline{p})|,$$

$$\sup_{\theta \leq 0} \theta r - M_{\pi, h}(\theta) \leq \sup_{\theta \leq 0} \left[ \theta (r - \underline{h}) - \log(\underline{p}) \right] \leq |\log(\underline{p})|.$$

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

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

**Lemma 2.6.** Suppose that Assumption (A1) holds. Then, the set $R(f, c) \subset \mathbb{R}^{n+1}$ defined in (29) is convex and compact.

**Proof.** Convexity is obvious. To establish that $R(f, c)$ is closed, consider any convergent sequence of vectors $\{\lambda^j\} \subset R(f, c)$. By the definition (29) we have,

$$\lambda^j \cdot c = 1, \quad \text{and} \quad \lambda^j \cdot f(x) \geq 0 \text{ for each } x \in X.$$

Obviously, these properties will be inherited by the limit. Thus $\lambda \in R(f, c)$, and hence the set $R(f, c)$ is closed.

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

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

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

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

$$\lambda^T(1, c_1, \ldots, c_n)^T + \epsilon e^i) \geq 0,$$

and since $\lambda^T c = 1$ it then follows that $1 + \epsilon \lambda_i \geq 0$. Similarly we get the bound $1 - \epsilon \lambda_i \geq 0$. Repeating this argument for each $i = 1, \ldots, n$, we can infer that

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

Since $\lambda^T c = 1$ and $c_0 = 1$, the above bounds imply upper and lower bounds on $\lambda_0$ as well. Hence $R(f, c)$ is closed and bounded, hence compact, as claimed. \qed

16
The following version of Cramér’s Theorem is used repeatedly below.

**Theorem 2.7.** Suppose that \( X \) is i.i.d. with one dimensional distribution \( \pi \) on \( \mathcal{B} \). Fix \( h \in C(X) \), 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:

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

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

\[
I_{\pi,h}(r) = \sup_{\theta \geq 0} \{ \theta r - M_{\pi,h}(\theta) \} = \inf \{ D(\mu \mid \pi) : \mu \text{ s.t. } \langle \mu, h \rangle \geq r \}
\]

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

\[
\frac{d\mu^*}{d\pi} = \frac{\exp(\theta^* h)}{\langle \pi, \exp(\theta^* h) \rangle}, \quad \text{and} \quad \langle \mu^*, h \rangle = r.
\]

**Proof.** Parts (i) and (ii) follow from Theorem 1.3 and the Contraction Principle. Part (iii) follows from [28, Theorem 1.5].

Define the functional \( K : \mathcal{M} \to \mathbb{R} \) by,

\[
K(\pi) := \inf \{ D(\mu \mid \pi) : \mu \in \mathcal{H}^1 \cup \mathcal{H} \}, \quad \pi \in \mathbb{P}.
\]

(34)

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

**Lemma 2.8.** For each \( \pi \in \mathcal{M} \) we have,

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

(35)

**Proof.** This follows directly from Theorem 2.7 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 (35) holds for all \( \pi \in \mathcal{M} \).

The following simple result is also used in an analysis of the functional \( K \).

**Lemma 2.9.** 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

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

(36)

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

For a proof of Proposition 2.10 see [29, Problem 7, page 236], following [29, Theorem 1, page 224].

**Proposition 2.10.** Suppose that $Y$ is finite-dimensional, and that the following two conditions hold:

(a) The optimal value $\kappa_0$ is finite, where

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

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

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

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

\[\square\]

The following result is required in an analysis of the linear program (7).

**Theorem 2.11.** (Lack of duality gap) Under Assumption (A1), for any $g \in C(X)$,

$$\max \{ \langle \pi, g \rangle : \pi \in M_1 \} = \min \{ \lambda^T c : \lambda \in \mathbb{R}^{n+1} \text{ s.t. } \lambda^T f \geq g \}. \tag{36}$$

Any distribution $\pi^*$ and vector $\lambda^*$ optimizing the respective linear programs (2) and (31) are together called a dual pair. A necessary and sufficient condition for a given pair $(\pi, \lambda)$ to form a dual pair is the alignment condition,

$$\lambda^T f - g \geq 0 \quad \text{and} \quad \langle \pi, \lambda^T f - g \rangle = 0. \tag{37}$$

**Proof.** A proof is contained in [26, 27]. We provide here a sketch since similar ideas are used in the proof of the existence of an optimizer $\pi^*$ in Theorem 1.4.

The state space $X$ is separable, so that there exists a countable dense set $\{x_i : i = 1, \ldots \} \subset X$. For each $q \geq 1$ we consider the finite state space,

$$X_q := \{x_i : 1 \leq i \leq q \} \tag{38}$$

Assumption (A1) implies that the set $\mathbb{P}$ is non-empty for all $q \geq 1$ sufficiently large. For such $q$, it is clear that the conclusions of Theorem 2.11 hold when $X$ is replaced by this finite state space.

Letting $(\pi^q, \lambda^q)$ denote the dual pair obtained for this optimization problem, and letting $v^q$ denote the optimal value, standard compactness arguments imply the existence of weak*-limits as $q \to \infty$. Any limiting pair is a solution to (36) since $v^q \uparrow v = \max \{ \langle \pi, g \rangle : \pi \in M_1 \}$.  \[\square\]
Proof of Theorem 1.4

Proof of (i): This is based on Proposition 2.10 with the identification,

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

We now verify the two required assumptions in Proposition 2.10: (a) the infimum (16) that defines \( L \) must be finite, and (b) the constraint vector \( c \) must lie in the interior of the set \( \Delta \). The first property is established in Lemma 2.5 (i), and the second is guaranteed by Assumption (A1).

Consequently, Proposition 2.10 implies the following expression for the worst-case rate function:

\[ L(\mu) = \inf_{\pi \in \pi} D(\mu \parallel \pi) = \max_{\lambda \in \mathbb{R}^n} \Psi(\lambda), \quad \text{where} \quad \Psi(\lambda) := \inf_{\pi \in \pi} \left\{ D(\mu \parallel \pi) + \lambda^T(\langle \pi, f \rangle - c) \right\}. \]

We now obtain an expression for \( \Psi \). Consider \( \lambda \in \mathbb{R}(f) \) satisfying \( \lambda^T f > 0 \) a.e. \([\mu], \) and define the positive measure \( \pi^\lambda \) through \( d\pi^\lambda = \frac{1}{\lambda^T f} \) (note that this may not be a probability measure.) Then we have

\[ D(\mu \parallel \pi^\lambda) + \lambda^T(\langle \pi^\lambda, f \rangle - c) = \langle \mu, \log \lambda^T f \rangle + 1 - \lambda^T c. \]

In order to show that the above expression is equal to the \( \Psi(\lambda) \), we must prove that \( \pi^\lambda \) achieves the minimum in the definition of \( \Psi(\lambda) \). Define \( \pi^\varrho := \varrho \pi + (1 - \varrho)\pi^\lambda \) for a given \( \pi \in \pi \), and \( \varrho \in [0, 1] \). Then it is straightforward to verify that

\[ \left. \frac{d}{d\varrho} \left( D(\mu \parallel \pi^\varrho) + \lambda^T(\langle \pi^\varrho, f \rangle - c) \right) \right|_{\varrho = 0} = \langle \pi^\lambda, \lambda^T f - \frac{d\mu}{d\pi^\lambda} F_A \rangle - \langle \pi, \lambda^T f - \frac{d\mu}{d\pi^\lambda} F_A \rangle, \]

where \( A \) is the support of \( \pi^\lambda \). From the definition of \( \pi^\lambda \), we have \( \lambda^T f F_A = \frac{d\mu}{d\pi^\lambda} F_A \), and therefore

\[ \left. \frac{d}{d\varrho} \left( D(\mu \parallel \pi^\varrho) + \lambda^T(\langle \pi^\varrho, f \rangle - c) \right) \right|_{\varrho = 0} = \langle \pi, (\lambda^T f) F_A \rangle. \]

Since \( \lambda^T f \geq 0 \), we have

\[ \left. \frac{d}{d\varrho} \left( D(\mu \parallel \pi^\varrho) + \lambda^T(\langle \pi^\varrho, f \rangle - c) \right) \right|_{\varrho = 0} \geq 0. \]

Since \( \pi \in \pi \) is arbitrary and \( D(\mu \parallel \cdot) \) is a convex function, we conclude that \( \pi^\lambda \) achieves the minimum in the expression for \( \Psi(\lambda) \). The formula for \( \Psi(\lambda) \) when \( \lambda \in \mathbb{R}(f) \) follows as a result.

We now finish the proof of the duality relation by showing that \( \Psi(\lambda) = -\infty \) whenever \( \lambda \notin R(f) \) or \( \lambda^T f \neq 0, \mu\text{-a.e.} \). Indeed, if \( \lambda \notin R(f) \) then \( \lambda^T f(x) < 0 \) for some \( x_0 \in X \). Let \( \pi^n := \mu + n\delta_{x_0} \), where \( \delta_{x_0} \) is the atom at \( x_0 \). Then \( D(\mu \parallel \pi^n) = 0 \) whereas \( \langle \pi^n, \lambda^T f \rangle \downarrow -\infty \).

In the latter case, in which \( \lambda^T f \) is not strictly positive a.e. \([\mu], \) it follows that the set \( A := \{ x : \lambda^T f(x) \leq 0 \} \) has positive \( \mu \)-measure. Consider the sequence of positive measures \( \pi^n \) defined through \( \frac{d\pi^n}{d\mu} = 1 + nF_A \). Clearly we have \( D(\mu \parallel \pi^n) + \langle \pi^n, \lambda^T f \rangle \downarrow -\infty \).
We conclude that \( d\mu \) is compact, the infimum is achieved by some \( \pi \) where

\[
\text{Proposition 2.10.}
\]

\[
\text{supremum in (i) are achieved. The fact that the supremum is achieved follows directly from Proposition 2.10.}
\]

We first restrict to the finite state space \( X_q \) defined in (38). Consider \( \mu \in \mathcal{M}_1(X_q) \), \( \pi \in \mathcal{M}_1(X) \), with \( \mu \prec \pi \). Then, the divergence is expressed,

\[
D(\mu \| \pi) = \sum_{i=1}^{q} \mu_i \log \left( \frac{\mu_i}{\pi_i} \right),
\]

where \( \mu_i := \mu(\{x_i\}) \), and \( \pi_i := \pi(\{x_i\}) \), and the worst-case rate-function is given by,

\[
L(\mu) = \inf_{\pi \in \mathcal{P}} \left\{ \sum_{i=1}^{q} \mu_i \log \left( \frac{\mu_i}{\pi_i} \right) \right\}.
\]

Observe that the infimum is taken over all of \( \mathcal{P} \subset \mathcal{M}_1 \).

The function \( \sum_{i=1}^{q} \mu_i \log \left( \frac{\mu_i}{\pi_i} \right) \) is lower semi-continuous as a function of \( \pi \in \mathcal{P} \). Since \( \mathcal{P} \) is compact, the infimum is achieved by some \( \pi^* \in \mathcal{P} \).

Let \( \lambda \in R(f, c) \) achieve the supremum in the duality relation for \( L(\mu) \). Proposition 2.10 implies that we must have \( \frac{d\mu}{d\pi} = \lambda^T f \) and \( \lambda^T c = 1 \).

This construction can be performed for any \( q \). Now, for arbitrary \( \mu \in \mathcal{M}_1 \), not necessarily discrete, there exists a sequence of probability distributions \( \{\mu^k\} \), each with finite support on \( \{x_1, x_2, \ldots\} \), such that \( \mu^k \xrightarrow{w} \mu \) as \( k \to \infty \), where the convergence is in the weak*-topology.

Since the infimum and supremum in the definition of \( L(\mu^k) \) is achieved for each \( \mu^k \), we have a sequence of vectors \( \{\lambda^k\} \subset R(f, c) \), and a sequence of probability measures \( \{\pi^k\} \subset \mathcal{P} \) such that \( \frac{d\mu^k}{d\pi^k} = \lambda^k \pi f \) and \( L(\mu^k) = D(\mu^k \| \pi^k) \) for each \( k \). The set \( \mathcal{P} \subset \mathcal{M}_1 \) is compact, and also \( R(f, c) \subset \mathbb{R}^{n+1} \) is compact, by Lemma 2.6. Thus there exists a sub-sequence \( \{k_i\} \) such that \( \lambda^{k_i} \to \lambda \) and \( \pi^{k_i} \to \pi^* \), with \( \lambda \in R(f, c) \) and \( \pi^* \in \mathcal{P} \). We claim that \( \frac{d\mu}{d\pi} = \lambda^T f \).

Since \( f : X \to \mathbb{R}^n \) is uniformly bounded, the convergence \( \lambda^{k_i} \pi^k \to \lambda^T f \) is uniform on \( X \).

Moreover, since \( \mu^{k_i} \xrightarrow{w} \mu \) and \( \pi^{k_i} \xrightarrow{w} \pi^* \), we have as \( i \to \infty \),

\[
\langle \mu^{k_i}, g \rangle \to \langle \mu, g \rangle \\
\text{and} \quad \langle \mu^{k_i}, g \rangle = \langle \pi^{k_i}, g \lambda^{k_i} \pi f \rangle \to \langle \pi^*, g \lambda^T f \rangle, \quad g \in C(X).
\]

We conclude that \( \frac{d\mu}{d\pi} = \lambda^T f \) as claimed.

We then have, from the duality relation for \( L \),

\[
D(\mu \| \pi^*) \geq L(\mu) \geq \langle \mu, \log \lambda^T f \rangle = D(\mu \| \pi^*).
\]

Thus, the probability measure \( \pi^* \) achieves the infimum in the definition of \( \mu \).

Proof of (ii) and (iii): The fact that \( L \) is convex follows directly from its formulation as a supremum of linear functionals in part (i). The finiteness of \( L \) is proved in Lemma 2.5 (i).

To show that \( L : \mathcal{M}_1 \to \mathbb{R}_+ \) is continuous, consider any convergent sequence of probability measures, \( \mu^k \xrightarrow{w} \mu \). From part (i) we know that there exist \( \{\lambda^k\} \subset R(f, c) \) and \( \{\pi^k\} \subset \mathcal{P} \) such that \( \frac{d\mu^k}{d\pi^k} = \lambda^k \pi f \), and \( L(\mu^k) = D(\mu^k \| \pi^k) \) 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 2.6), we can construct if necessary a further subsequence so that \( \lambda^{k_i} \to \lambda \) and \( \pi^{k_i} \xrightarrow{w} \pi \), for some \( \lambda \in R(f,c) \), and \( \pi \in \mathbb{P} \). As in the proof of part (i), it follows that \( \frac{d\lambda}{d\pi} = \lambda^T f \) and \( L(\mu) = D(\mu \| \pi) \). Since \( f \) is a bounded function, we must have \( \lambda^{k_i} r \to \lambda^T f \) uniformly on \( X \).

Consider now the functions \( \{(\lambda^{k_i} r f) \log \lambda^{k_i} r f : i \geq 1\} \). Since \( x \log x \) is a continuous function on \( \mathbb{R} \), it follows that
\[
(\lambda^{k_i} r f) \log \lambda^{k_i} r f \to (\lambda^T f) \log \lambda^T f, \quad \text{uniformly on } X,
\]
and, since \( \pi^{k_i} \xrightarrow{w} \pi^* \), we have
\[
\langle \pi^{k_i}, (\lambda^{k_i} r f) \log \lambda^{k_i} r f \rangle \to \langle \pi, (\lambda^T f) \log \lambda^T f \rangle.
\]
From the identities,
\[
L(\mu^{k_i}) = \langle \pi^{k_i}, (\lambda^{k_i} r f) \log \lambda^{k_i} r f \rangle \quad \text{and} \quad L(\mu) = \langle \pi, (\lambda^T f) \log \lambda^T f \rangle,
\]
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,
\[
\mathcal{Q}_\beta(\mathbb{P}) = \{\mu: L(\mu) < \beta\}.
\]
This set is convex and open since the functional \( L \) is convex and continuous.

In part (i) we showed that the infimum in (16) is achieved by some \( \pi \in \mathbb{P} \), from which it follows that
\[
\mathcal{Q}_\beta^+(\mathbb{P}) = \{\mu: L(\mu) \leq \beta\}.
\]
Since \( L \) is convex and continuous, \( \mathcal{Q}_\beta^+(\mathbb{P}) \) is convex and closed. In fact, \( \mathcal{Q}_\beta^+(\mathbb{P}) \) is compact since \( X \) and hence \( \mathcal{M}_1 \) is compact.

In order to identify the closure of \( \mathcal{Q}_\beta(\mathbb{P}) \), consider any \( \mu \in \mathcal{Q}_\beta^+(\mathbb{P}) \) such that \( \mu \notin \mathcal{Q}_\beta(\mathbb{P}) \). From part (i), we know that there exists \( \pi \in \mathbb{P} \) with \( L(\mu) = D(\mu \| \pi) = \beta \). On writing \( \mu^\epsilon := \epsilon \pi + (1 - \epsilon)\mu \), we have by convexity of \( L \),
\[
L(\mu^\epsilon) \leq \epsilon L(\pi) + (1 - \epsilon)L(\mu) < \beta, \quad \epsilon \in (0,1).
\]
Moreover, \( \mu^\epsilon \xrightarrow{w} \mu \) as \( \epsilon \downarrow 0 \), and from continuity of \( L \) it follows that \( L(\mu^\epsilon) \to L(\mu) = \beta \) as \( \epsilon \downarrow 0 \). Thus \( \mu \) lies in the closure of \( \mathcal{Q}_\beta(\mathbb{P}) \). Since \( \mathcal{Q}_\beta^+(\mathbb{P}) \) is closed, it follows that closure \( \mathcal{Q}_\beta(\mathbb{P}) = \mathcal{Q}_\beta^+(\mathbb{P}) \). \( \square \)

**Lemma 2.12.** Let \( \mathbb{P} \) be a moment class that satisfies Assumption (A1), and suppose that \( \pi^* \in \mathbb{P} \) is \((h,r,+)\)-extremal for some \( r \in (\overline{r},h] \). Then, there exists \( \theta^* < \infty \) such that
\[
L_h(r) = I_{\pi^*,h}(r) = \left[ \theta^* r - M_{\pi^*,h}(\theta^*) \right] = \max_{\theta \geq 0} \left[ \theta r - M_{\pi^*,h}(\theta) \right].
\]
Proof. The existence of $\pi^*$ satisfying $I_{\pi^*, h}(r) = I_{\pi^*, h}(r)$ is immediate from Theorem 1.4 and (15). Hence, it remains to show that the supremum in the definition of $I_{\pi^*, h}(r)$ is achieved at some finite $\theta^*$.

We prove this by contradiction: Suppose that no finite $\theta^* \in \mathbb{R}_+$ exists, so that

$$I_{\pi^*, h}(r) > \left[ \theta r - M_{\pi^*, h}(\theta) \right], \quad \theta < \infty$$

and

$$I_{\pi^*, h}(r) = \lim_{\theta \to \infty} \left[ \theta r - M_{\pi^*, h}(\theta) \right].$$

(39)

Fix $s \in (r, h)$, and let $\nu^* \in \mathbb{P}$ denote a corresponding extremal distribution (whose existence again follows from Theorem 1.4 and (15).) The probability distributions \{\pi^*, \nu^*\} satisfy,

$$I_{\pi^*, h}(r) = I_{\nu^*, h}(s) = \sup_{\theta \geq 0} \left[ \theta s - M_{\nu^*, h}(\theta) \right].$$

Lemma 2.5 (ii) implies that $I_{\pi^*, h}$ is finite at $r$ and $s$.

For $\varrho \in (0, 1)$, define $\pi^* := \varrho \nu^* + (1 - \varrho) \pi^*$. We then have, from concavity of the logarithm,

$$M_{\pi^* \varrho, h}(\theta) \geq \varrho M_{\nu^*, h}(\theta) + (1 - \varrho) M_{\pi^*, h}(\theta), \quad \theta \geq 0,$$

and this implies the following upper bounds:

$$\theta r - M_{\pi^* \varrho, h}(\theta) \leq \varrho \left[ \theta s - M_{\nu^*, h}(\theta) \right] + (1 - \varrho) \left[ \theta r - M_{\pi^*, h}(\theta) \right] - \varrho \theta (s - r)$$

$$\leq \varrho L_{h}(s) + (1 - \varrho) L_{h}(r) - \varrho \theta (s - r)$$

$$= L_{h}(r) + \varrho \left[ - \theta (s - r) + L_{h}(s) - L_{h}(r) \right], \quad \theta \geq 0.$$

Define $\theta_0 = (s - r)^{-1} [L_{h}(s) - L_{h}(r) + 1]$. We then have,

$$\theta r - M_{\pi^* \varrho, h}(\theta) \leq L_{h}(r) - \varrho, \quad \theta \geq \theta_0.$$

Under (39) it follows that the following strict inequality holds,

$$\theta r - M_{\pi^*, h}(\theta) < L_{h}(r), \quad 0 \leq \theta \leq \theta_0.$$

Hence, for all $\varrho > 0$ sufficiently small we have the identical inequality with $\pi^0$: for some $\delta > 0$,

$$\theta r - M_{\pi^* \varrho, h}(\theta) \leq L_{h}(r) - \delta, \quad 0 \leq \theta \leq \theta_0.$$

For such $\varrho$ we conclude that,

$$I_{\pi^* \varrho, h}(r) := \sup_{\theta \geq 0} \left[ \theta r - M_{\pi^* \varrho, h}(\theta) \right] \leq L_{h}(r) - \min(\delta, \varrho).$$

This contradicts minimality of $L_{h}(r)$, and completes the proof. \qed
Proof of Theorem 2.1

Part (i) (Necessity): Recall the definition of the functional $K$ given in (34). As illustrated in Figure 1, the inclusion $Q_\beta(\mathbb{P}) \subset H^0$ holds, and consequently $D(\mu \parallel \pi) \geq \beta$ for every $\mu \in H^1 \cup H$, and $K(\pi) \geq \beta$ for every $\pi \in \mathbb{P}$.

By assumption $\mu^* \in H$ and $D(\mu^* \parallel \pi^*) = \beta$, so the measure $\pi^*$ achieves this lower bound:

$$K(\pi^*) = \inf_{\pi \in \mathbb{P}} K(\pi) = \beta. \quad (40)$$

Lemma 2.12 states that the supremum over $\theta$ in (35) is attained by some $\theta^* \geq 0$ when $\theta = \theta^*$. Moreover, Theorem 2.7 implies the identity,

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

and we must then have $\theta^* > 0$ since $K(\pi^*) = \beta > 0$, and

It is straightforward to verify that the infimum in (40) meets the conditions of Proposition 2.10, from which we obtain,

**Lemma 2.13.** There exists $\lambda \in \mathbb{R}^{n+1}$ such that

$$K(\pi^*) = K(\pi^*) + \lambda^T (\langle \pi^*, f \rangle - c) = \inf_{\pi \in \mathcal{M}} \{ K(\pi) + \lambda^T (\langle \pi, f \rangle - c) \}. \quad (41)$$

$\square$

We apply Lemma 2.13 to establish the alignment condition (28): Consider any $\pi^0 \in \mathcal{M}$ such that the supremum in (35) is achieved for some $\theta^0 \in \mathbb{R}_+$. Lemma 2.9 then implies the following bound,

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

where $g_0 := g_{\pi^0, h}$ as defined in Lemma 2.9. Consequently,

$$K(\pi^*) + \lambda^T (\langle \pi^*, f \rangle - c) \geq K(\pi^0) + \lambda^T (\langle \pi^0, f \rangle - c) + \langle \pi^* - \pi^0, \lambda^T (f - c) - g_0 \rangle,$$

and on combining this with (41) we obtain,

$$\langle \pi^* - \pi^0, \lambda^T (f - c) - g_0 \rangle \leq 0.$$

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

$$g_0(x) \leq \lambda^T (f(x) - c).$$

It is clear that $M_{\pi^0, h} \to M_{\pi^*, h}$ as $\epsilon \downarrow 0$, uniformly for $\theta$ 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

$$\frac{e^{\theta^* h(x)}}{\langle \pi, e^{\theta^* h} \rangle} \leq \lambda^T (f(x) - c), \quad x \in X.$$
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 $\langle \pi^*, \lambda^T (f - c) - g_\ast \rangle \leq 0$. The representation (28) follows on taking logarithms.

Part (ii) (Sufficiency): Recall that (35) gives,

$$K(\pi^*) = \sup_{\theta \geq 0} \{ \theta r - \log \langle \pi^*, \exp(\theta h) \rangle \}.$$

Since $\frac{d\pi^*}{d\pi} = \frac{\exp(\theta^* h)}{\exp(\theta h)}$ and $\langle \mu^*, h \rangle = r$, we conclude from Theorem 2.7 that the above supremum is achieved at $\theta = \theta^*$, so that

$$K(\pi^*) = \theta^* r - \log \langle \pi^*, \exp(\theta^* h) \rangle = \theta^* r - \log \langle \pi^*, b_0 \lambda^T f \rangle = \theta^* r - \log (b_0 \lambda^T c).$$

Moreover, for any $\pi \in \mathbb{P}$,

$$K(\pi) \geq \theta^* r - \log \langle \pi, \exp(\theta^* h) \rangle \geq \theta^* r - \log \langle \pi, b_0 \lambda^T f \rangle = \theta^* r - \log (b_0 \lambda^T c),$$

where the second inequality follows from the assumed condition (28). Consequently we have $K(\pi) \geq K(\pi^*)$ for any $\pi \in \mathbb{P}$.

It follows from the definition of $K(\pi)$ that if $\mu$ satisfies $\langle \mu, h \rangle \leq r$ then for every $\pi \in \mathbb{P}$ we must have

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

In fact we must have $D(\mu \| \pi) > \beta$ if $\langle \mu, h \rangle > r$, since $K(\pi)$ is always achieved by a distribution $\mu^*$ with $\langle \mu^*, h \rangle = r$ from Theorem 2.7. Consequently we must have $Q_\beta(\mathbb{P}) \subset \mathcal{H}^1$, which implies $Q_\beta(\mathbb{P}) \subset \mathcal{H} \cap \mathcal{H}^1$, since $Q_\beta(\mathbb{P})$ is the closure of $Q_\beta(\mathbb{P})$ from Theorem 1.4. Since $\mu^* \in \mathcal{H} \cap Q_\beta(\mathbb{P})$, the hyperplane $\mathcal{H}$ must be a supporting hyperplane for $Q_\beta(\mathbb{P})$. \hfill \Box

The two lemmas below are consequences of Theorem 2.1. 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 2.14.** 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

$$\exp(\theta^* h) \begin{cases}
= b_0 \lambda^T f = b_0 \frac{d\mu^*}{d\pi}, & \text{a.e. } [\pi^*] \\
\leq b_0 \lambda^T f & \text{everywhere.}
\end{cases} \tag{42}$$

Proof. The fact that $\mathcal{H}$ is a supporting hyperplane for $Q_\beta(\mathbb{P})$, together with Theorem 2.1, 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 (42) is a necessary condition for $\pi^*$ to be an extremal distribution, where $b_0 := \exp(\theta^* r - \beta^*)$.

Conversely, if $\pi^*, \mu^*$ satisfy (42) along with $D(\mu^* \| \pi^*) = \beta^*$, then from Theorem 2.1, the hyperplane $\mathcal{H}$ supports $Q_\beta(\mathbb{P})$, and therefore the pair $\{\pi^*, \mu^*\}$ solves (26). Thus (42) is also a sufficient condition for $\pi^*$ to be an extremal distribution. \hfill \Box
Lemma 2.15. Let $\mathbb{P}$ be a moment class that satisfies Assumption (A1). Suppose that $\pi^* \in \mathbb{P}$ is $(h, r, +)$-extremal for some $r \in (\tau_h, \overline{h})$, and let $\theta^* > 0$ be the constant given in Theorem 2.14. Then, $\pi^*$ is also an optimizer of the infinite-dimensional linear program (7) that defines the worst-case moment-generating function, with $\theta = \theta^*$.

Proof. From (42) 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 $\pi^*$ solves (7). $\square$

Proof of Theorem 2.2

The proof follows from Lemma 2.15 and Lemma 2.12: From these results and maximality of $\overline{M}_h$ we have,
\[
L_h(r) \geq \theta r - M_{\pi^*, h}(\theta) \geq \theta r - \overline{M}_h(\theta), \quad \theta \geq 0,
\]
and all inequalities become equalities when $\theta = \theta^*$. $\square$
3 Conclusions & Future Directions

In this paper we have established an explicit formula for the worst-case large deviations rate-functions $L$ and $\{L_h : h \in C(X)\}$. The geometric structure of the divergence set $Q^+_{\beta}(P)$ plays a central role in interpreting the results, and is a valuable tool in analysis.

Potential directions for future research include,

(a) The results of this paper apply only to i.i.d. processes. It is likely that some of these conclusions can be extended to include Markov processes. Recent results on exact large deviations for general state space chains may be helpful [30, 31].

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

(c) Applications of Theorem 1.4 to robust hypothesis testing are considered in [32].

(d) Results from [21] and this paper provided inspiration for the research described in [25]. The discrete nature of extremal distributions provided motivation for algorithms computing 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.
References


