JOURNAL OF OPTIMIZATION THEORY AND APPLICATIONS: Vol. 58, No. 2, AUGUST 1988

The Role of Duality in Optimization Problems Involving Entropy Functionals with Applications to Information Theory[1]

A. BEN-TAL,[2] M. TEBOULLE,[3] AND A. CHARNES[4]

Communicated by M. Avriel

Abstract. We consider infinite-dimensional optimization problems involving entropy-type functionals in the objective function as well as as in the constraints. A duality theory is developed for such problems and applied to the reliability rate function problem in information theory.

Key Words. Optimization in infinite-dimensional spaces, duality in convex optimization, entropy, divergence, information theory, channel capacity, reliability rate function, error exponent function.

1. Introduction

Extremum problems involving entropy-type functionals appear in a diversity of applications. To mention just a few: statistical estimation and hypothesis testing [Kullback–Leibler (Ref. 1), Kullback (Ref. 2), Akaike (Ref. 3)], traffic engineering [Charnes *et al.* (Ref. 4)], marketing [Charnes *et al.* (Ref. 5)], accounting [Charnes and Cooper (Refs. 6, 7)], information theory [Shannon (Ref. 8)].

In the majority of these applications, the extremum problems involved are studied only for the case of finite distributions. Extensions to arbitrary distributions were derived recently by Ben-Tal and Charnes (Ref. 9). The

[1] This research was supported by ONR Contracts N00014-81-C-0236 and N00014-82-K-0295 with the Center for Cybernetics Studies, University of Texas, Austin, Texas. The first author was partly supported by NSF.

[2] Professor, Faculty of Industrial Engineering and Management, Technion, Israel Institute of Technology, Haifa, Israel, and University of Michigan, Ann Arbor, Michigan.

[3] Assistant Professor, Department of Mathematics and Statistics, University of Maryland, Baltimore County Campus, Baltimore, Maryland.

[4] Director, Center for Cybernetic Studies, University of Texas, Austin, Texas.

0022-3239/88/0800-0209$06.00/0

extremum problem is set up as an infinite-dimensional convex program with linear equality constraints, namely:

$$(\mathrm{A}) \quad \inf_{f \in D} \left\{ \int_T f(t) \log[f(t)/g(t)]\, dt : \right.$$

$$\left. \int_T f(t) a_i(t)\, dt = \theta_i, \qquad i = 1, \ldots, m \right\},$$

where D is the convex subset of density functions with support T and $g(\cdot)$ is a given density in D.

It is shown in Ref. 9 that the dual problem is the unconstrained finite-dimensional concave program:

$$(\mathrm{B}) \quad \sup_{y \in \mathbb{R}^m} \left\{ y^t \theta - \log \int_T g(t) \exp\left[\sum_{i=1}^m y_i a_i(t) \right] dt \right\}.$$

The dual pair (A), (B) has a very interesting statistical interpretation. Let $\{\theta_i\}_{i=1}^m$ be parameters of the distribution, estimated in terms of a sample $x = (x_1, \ldots, x_n)$ by

$$\hat{\theta}_i(x) = \hat{\theta}_i(x_1, \ldots, x_n) = (1/n)(a_i(x_1) + \cdots + a_i(x_n)),$$

and let these estimates replace θ_i in the constraints of (A). Consider now the problem of finding the maximum likelihood estimator $\pi^*(x)$ of the parameter vector $\pi = (\pi_1, \ldots, \pi_m)^t$ in the exponential family generated by the (fixed) density $g(t)$, i.e.,

$$f(t|\pi) = g(t) c(\pi) \exp\left[\sum_{i=1}^m \pi_i a_i(t) \right],$$

where $c(\pi)$ is a normalizing constant, i.e.,

$$c(\pi)^{-1} = \int_T g(t) \exp\left[\sum_{i=1}^m \pi_i a_i(t) \right] dt.$$

The likelihood function is

$$\prod_{j=1}^n f(x_j|\pi) = \left\{ \prod_{j=1}^n g(x_j) \right\} \cdot c(\pi)^n \exp\left[\sum_j \sum_i \pi_i a_i(x_j) \right];$$

hence,

$$(1/n) \log(\text{likelihood}) = \text{const} + \log c(\pi) \exp[\textstyle\sum \pi_i \hat{\theta}_i(x)];$$

therefore, the maximum likelihood estimator $\pi^*(x)$ is obtained by solving

$$\max_{\pi \in \mathbb{R}^m} \{\sum \pi_i \hat{\theta}_i(x) - \log c^{-1}(\pi)\}$$

$$= \max_{\pi \in \mathbb{R}^m} \left\{ \sum_{i=1}^{m} \pi_i \hat{\theta}_i(x) - \log \int_T g(t) \exp\left[\sum_{i=1}^{m} \pi_i a_i(t) \right] dt \right\}.$$

The latter is precisely the dual problem (B). Thus, for the exponential family, statistical information theory and maximum likelihood approach are dual principles.

Many problems in information theory, however, cannot be stated just with linear constraints as in problem (A); they contain also (nonlinear) entropy-type inequality constraints. It is the purpose of this paper to derive duality results for such problems and to demonstrate their power and elegance in treating such problems.

As a motivation we begin by describing the channel capacity problem of information theory. Consider a communication channel described by an input alphabet $A = \{1, \ldots, n\}$, an output alphabet $B = \{1, \ldots, m\}$, and a probability transition matrix $Q = \{Q(k|j)\}$, where $Q(k|j)$ is the probability of receiving the output letter $k \in B$ when input letter $j \in A$ is transmitted.

The capacity of the channel is defined as

$$C = \max_{p \in P^n} I(p, Q)$$

$$\triangleq \max_{p \in P^n} \sum_{k=1}^{m} \sum_{j=1}^{n} p_j Q(k|j) \log \left[Q(k|j) \Big/ \sum_{l=1}^{n} p_l Q(k|l) \right], \tag{1}$$

where

$$\mathbb{P}^n \triangleq \left\{ p \in \mathbb{R}^n : p_j \geqslant 0, \forall j, \sum_{j=1}^{n} p_j = 1 \right\} \tag{2}$$

is the set of all probability distributions on the channel input and $I(p, Q)$ is known as the average mutual information between the channel input and channel output. Channel capacity is the basic concept of Shannon's mathematical theory of communication (later called information theory). For more details on the notion of capacity and its significance, the reader is referred to Shannon (Ref. 8), Gallager (Ref. 10), and Jelineck (Ref. 11).

Roughly speaking, the basic theorem of information theory, the so-called noisy channel coding theorem, states that, if the channel has capacity C, it is possible to transmit over this channel messages of sufficiently large length at rate $R < C$ and still be able to decode them with a probability of error as small as desired. The upper bound on the probability of error is given in terms of an exponential decreasing function of the so-called reliability rate function $E(R)$. In the classical proof of the coding theorem, the function $E(R)$ is derived via a sequence of mathematical manipulations;

see, e.g., Gallager (Ref. 12) and Csiszar (Ref. 13). Blahut (Ref. 14) has enlightened many basic problems of coding theory be defining $E(R)$ as a saddle function problem, involving the Kullback–Leibler relative entropy functional, namely, for a given channel matrix $P(k|j)$,

$$E(R)=\max_{p\in\mathbb{P}^n}\min_{Q\in\mathcal{Q}(R)}\sum_{k=1}^{m}\sum_{j=1}^{n}p_jQ(k|j)\log[Q(k|j)/P(k|j)], \tag{3}$$

where

$$\mathcal{Q}(R)=\{Q\colon I(p,Q)\leqslant R\},\quad R \text{ a positive scalar.}$$

Starting from this definition, Blahut (Ref. 14) proved that $E(R)$ can be expressed by the conventional parametric form originally proposed by Gallager (Ref. 12), namely,

$$E(R)=\max_{\delta\geqslant 0}\max_{p\in\mathbb{P}^n}\left\{-\delta R-\log\sum_{k=1}^{m}\left\{\sum_{j=1}^{n}p_jP(k|j)^{1/(1+\delta)}\right\}^{1+\delta}\right\}. \tag{4}$$

A new proof of this result is given here in Section 3, via the duality theory developed in Section 2. The duality framework can be applied to a variety of other extremum problems of information theory; see, e.g., Blahut (Ref. 14), Table I, p. 417.

In particular, more than one entropy-type constraint can be easily dealt with, and the general (not necessarily discrete) distribution case can be considered.

2. Duality Theory for Linear and Entropy Constrained Programs

Let dt be a σ-finite additive measure defined on a σ-field of the subsets of a measurable space T, and let $L^1\triangleq L^1(T, dt)$ be the usual Lebesgue space of measurable, real-valued functions x on T so that

$$\|x\|\triangleq\int_T|x(t)|\,dt<\infty.$$

Let

$$\mathbb{D}=\left\{x\in L^1\colon x(t)\geqslant 0,\ \text{a.e.},\ \int_T x(t)\,dt=1\right\}$$

be the convex subset of L^1, i.e., $\mathbb{D}$ is the set of all probability densities $x(\cdot)$ on T.

Consider the infinite-dimensional optimization problem:

$$\text{(P)}\quad \inf\int_T x(t)\log[x(t)/c_0(t)]\,dt,$$

subject to

$$\int_T a_i(t)x(t)\,dt \geq b_i, \qquad i \in I \triangleq \{1, \ldots, m\}, \tag{5}$$

$$\int_T x(t)\log[x(t)/c_k(t)] \leq e_k, \qquad k \in K \triangleq \{1, \ldots, p\}, \tag{6}$$

and $x(t) \in \mathbb{D} \subset L^1$, where $c_k : T \to \mathbb{R}$, $k \in \{0\} \cup K$ are given summable positive functions, $a_i : T \to \mathbb{R}$ are given continuous functions, and $\{b_i\}_{i\in I}$, $\{e_k\}_{k\in K}$ are given real numbers.

Here and henceforth,

$$0 \log 0 = \lim_{t \to 0^+} t \log t = 0.$$

A dual representation of Problem (P) will be derived via Lagrangian duality. Recall that, for a convex optimization problem,

(A) $\inf\{f(x): g(x) \leq 0, x \in C \subset X\}$,

where $f : C \to \mathbb{R}$, $g : C \to \mathbb{R}^m$ are convex functions defined on a convex subset C of a linear space X, the Lagrangian for problem (A) is defined as $L : C \times \mathbb{R}^m_+ \to \mathbb{R}$, given by

$$L(x, y) = f(x) + y^t g(x).$$

The dual objective function is

$$h(y) = \inf_{x \in C} L(x, y),$$

and then the dual problem (B) associated with (A) is defined as

(B) $\sup_{y \geq 0} h(y)$.

The main result concerning the dual pair (A) and (B) is the existence of a saddle point (x^*, y^*) for L or, equivalently, the validity of a strong dualilty result,[5]

$$\inf(A) = \max(B).$$

Under the familiar Slater regularity condition,

(S) $\exists x \in C: g(x) < 0$,

the strong duality relation is guaranteed. More precisely, we have the following theorem [see, e.g., Rockafeller (Refs. 15, 16), Laurent (Ref. 17), and Poinstein (Ref. 18)].

[5] We follow the convention of writing min (max) if the infinum (supremum) is attained.

Theorem 2.1. Assume that $\inf(A) < \infty$ and that the regularity assumption (S) holds. Then,

$$\inf(A) = \max(B).$$

Remark 2.1. The regularity condition (S) is, in fact, related to the notion of a stably set problem. More details are available in Rockafeller (Ref. 15) and Laurent (Ref. 17. especially Theorem 7.6.1, p. 403).

Remark 2.2. A result of the type of Theorem 2.1 has typically a symmetric version; i.e., if (B) is assumed stably set, then $\min(A) = \sup(B)$; see Rockafeller (Ref. 15, Theorem 4, p. 179).

We now return to the primal entropy problem (P). The derivation of its dual objective function is based on the following simple result.

Lemma 2.1. Let $s(t)$ be a given positive summable function,

$$\int_T s(t)\, dt = S < \infty.$$

Then,

$$\min_{x \in \mathbb{D}} \int_T x(t) \log[x(t)/s(t)]\, dt = -\log S,$$

where the optimal probability density is

$$x^*(t) = s(t)/S, \quad \text{a.e.}$$

Proof. Define $h(t) = s(t)/S$. Then, $h(t) \in \mathbb{D}$; hence, by Theorem 3.1, p. 14, Ref. 2, we have

$$\inf_{x \in \mathbb{D}} \int_T x(t) \log[x(t)/h(t)]\, dt = 0,$$

where the infimum is attained for

$$x^*(t) = h(t) = s(t)/S.$$

Then, using the identity

$$\int_T f(t) \log[f(t)/s(t)]\, dt = \int_T f(t) \log[f(t)/h(t)] - \log \int_T s(t)\, dt,$$

the result follows. □

The Lagrangian for problem (P) is $L: \mathbb{D} \times \mathbb{R}_+^m \times \mathbb{R}_+^p \to \mathbb{R}$,

$$L(x, y, \lambda) = b^t y - e^t \lambda + \int_T \left\{ \log[x(t)/c_0(t)] - \sum_{i \in I} y_i a_i(t) + \sum_{k \in K} \lambda_k \log[x(t)/c_k(t)] \right\} x(t)\, dt, \tag{7}$$

and thus the dual problem (D) associated with (P) is defined as

$$\sup\{\inf_{x \in D} L(x, y, \lambda) : y \in \mathbb{R}_+^m, \lambda \in \mathbb{R}_+^p\}.$$

The next result shows that the dual problem (D) can be expressed simply as a finite-dimensional concave program involving only nonnegative constraints.

Theorem 2.2. The dual problem of (P) is given by

$$(\mathrm{D}) \quad \sup_{\substack{y \in \mathbb{R}_+^m \\ \lambda \in \mathbb{R}_+^p}} \left\{ y^t b - \lambda^t e - \rho \log \int_T c_0(t) \exp[(1/\rho)\lambda^t B(t) + y^t A(t)]\, dt \right\},$$

where

$$\rho = 1 + \sum_{k=1}^{p} \lambda_k,$$

$$A(t) = (a_1(t), \ldots, a_m(t))^t,$$

$$B(t) = (B_1(t), \ldots, B_p(t))^t,$$

$$B_k(t) = \log[c_k(t)/c_0(t)], \qquad \forall k \in K = \{1, \ldots, p\}.$$

Proof. The Lagrangian defined in (7) can be written, after some algebraic manipulations, as

$$L(x, y, \lambda) = -y^t b - \lambda^t e + \int_T x(t) \log\left\{ x(t)^{[1+\sum_{k=1}^p \lambda_k]} \Big/ \left[\prod_{k=0}^{p} c_k(t)^{\lambda_k} \right] \exp[y^t A(t)] \right\} dt.$$

Then, defining

$$\rho = 1 + \sum_{k=1}^{p} \lambda_k,$$

$$B_k(t) = \log[c_k(t)/c_0(t)],$$

a little algebra shows that the dual objective function can be expressed as

$$h(y, \lambda) = y^t b - \lambda^t e + \rho \inf_{x \in \mathbb{D}} \int_T x(t) \log\{x(t)/c_0(t) \exp[(1/\rho)(\lambda^t B(t) + y^t A(t)]\}\, dt.$$

Now, applying Lemma 2.1 with

$$s(t) = c_0(t) \exp[(1/\rho)(\lambda^t B(t) + y^t A(t)],$$

we get the desired result. □

Duality results for the pair of problems (P)-(D) will now follow by setting problem (P) as a convex program of the type (A) and then applying Theorem 2.1.

Theorem 2.3. (a) If (P) is feasible, then inf(P) is attained and min(P) = sup(D). Moreover, if there exists $x \in \mathbb{D}$ satisfying the constraints (5), (6) strictly, then sup(D) is attained and min(P) = max(D).

(b) If $x^* \in \mathbb{D}$ solves (P) and $y^* \in \mathbb{R}^m_+$, $\lambda^* \in \mathbb{R}^p_+$ solves (D), then

$$x^*(t) = \frac{c_0(t) \exp[(1/\rho)(\lambda^{*t} B(t) + y^{*t} A(t))]}{\int_T c_0(t) \exp[(1/\rho)(\lambda^{*t} B(t) + y^{*t} A(t))]\, dt}, \qquad \text{a.e.}$$

Proof. In order to apply Theorem 2.1, we need to set problem (P) in the format of the convex program (A). Thus, consider the linear operator $A: L^1 \to \mathbb{R}^m$, given by

$$x \to \begin{bmatrix} \int_T a_1(t)x(t)\, dt \\ \vdots \\ \int_T a_m(t)x(t)\, dt \end{bmatrix};$$

and, for $k \in \{0\} \cup K$, define the integral functionals

$$I_k(x) = \begin{cases} \int_T x(t) \log[x(t)/c_k(t)]\, dt, & \text{if } x \in \mathbb{D}, \\ \infty, & \text{otherwise.} \end{cases}$$

Then problem (P) can be written as a convex optimization problem,

(P) $\inf\{I_0(x): Ax \geq b, I_k(x) \leq e_k, k \in K, x \in \mathbb{D}\}$.

Note that (P) corresponds to (A) with

$$X := L^1, \qquad C := \mathbb{D}, \qquad f(x) := I_0(x), \qquad g(x) := \begin{bmatrix} b - Ax \\ I_1(x) - e_1 \\ \vdots \\ I_p(x) - e_p \end{bmatrix},$$

and then the results follow from Theorem 2.1. In fact, since the dual (D), given in Theorem 2.2, has only nonnegative constraints $y \geqslant 0$, $\lambda \leqslant 0$, it satisfies the strongest constraint qualification, implying by Remark 2.2 lack of duality gap and attainment of the primal infimum. Thus, the first part of conclusion (a) follows. The second part follows directly from Theorem 2.1 itself. Moreover, part (a) implies the existence of a saddle point $(x^*(t), y^*, \lambda^*) \in \mathbb{D} \times \mathbb{R}_+^m \times \mathbb{R}_+^p$, so

$$\min_{x \in \mathbb{D}} L(x, y^*, \lambda^*) = L(x^*, y^*, \lambda^*),$$

and the expression for x^* given in (b) follows from the last part of Lemma 2.1. □

3. An Application in Information Theory

In this section, we apply the duality relation for problem (P) to treat in a unified simple way the reliability rate function problem described in the introduction. While the results developed in Section 2 are applicable to the case of general probability distributions, we restrict ourselves here to the case of finite discrete probability distributions, since they include most of the interesting problems appearing in information theory. We begin with some further notations and definitions, following closely the terminology of Ref. 14.

The relative entropy or discrimination between two discrete (finite) distributions p, q playing a fundamental role in statistical information theory is a function $J: \mathbb{P}^n \times \mathbb{P}^n \to \mathbb{R}$, defined by

$$J(p, q) = \sum_{k=1}^{n} p_k \log(p_k / q_k). \tag{8}$$

It is well known that J is convex in each of its arguments, nonnegative, and equal to zero if and only if $p_k = q_k$, $\forall k$; see, e.g., Ref. 2.

Similarly, one defines the average discrimination by

$$J(Q, P) = \sum_{j=1}^{n} \sum_{k=1}^{m} p_j Q(k|j) \log[Q(k|j)/P(k|j)], \tag{9}$$

where p, Q, P are as defined in the introduction.

In the rest of this paper, we simplify the notations: probability transition matrices, like $P(k|j)$, are denoted P_{kj} and summation indices are dropped.

An error exponent function is defined in Blahut (Ref. 14) as the following (single) entropy-constrained program:

(E) $e(r) = \min\{J(q, q_2): q \in \mathbb{P}(r)\}$,

where

$$\mathbb{P}(r) = \{q \in \mathbb{P}^n : J(q, q_1) \leqslant r\}.$$

r is a given positive scalar and q_1, q_2 are given distributions in $\mathbb{P}^n$. Problem (E) just defined is a special case of problem (P), described in Section 2, with $I = \varnothing$ (i.e., no linear constraints), $K = \{1\}$, and $c_0(t)$, $c_1(t)$ corresponding here to the discrete finite distributions q_2, q_1 respectively. Moreover, since problem (E) consists of minimizing continuous functions over the compact set $\mathbb{P}(r)$, the minimum is attained; we know also from Theorem 2.2 that the dual problem (H) corresponding to (E) inolves only nonnegative constraints, hence satisfying the strongest constraint qualifications.

According to Theorem 2.2 and Theorem 2.3, by setting

$$\rho = 1 + \lambda_1 = 1 + \delta, \qquad e_1 = r,$$

we get the following theorem.

Theorem 3.1. A dual representation of (E) is the program

$$\text{(H)} \quad e(r) = \max_{\delta \geqslant 0} \{-\delta r - \log[\textstyle\sum q_{1k}^{\delta/(1+\delta)} q_{2k}^{\delta/(1+\delta)}]^{1+\delta}\}.$$

Moreover, if $q^* \in \mathbb{P}^n$ solves (E) and $\delta^* \geqslant 0$ solves (H), then

$$q_k^* = \frac{q_{1k}^{\delta/(1+\delta)} q_{2k}^{1/(1+\delta)}}{\sum\limits_k q_{1k}^{\delta/(1+\delta)} q_{2k}^{1/(1+\delta)}}.$$

We recover here a result obtained in Ref. 14, Theorem 7.

We now derive the dual representation of $E(R)$ by reference to the error exponent function $e(r)$. Recalling the definition of the reliability rate function given in the introduction [see Eq. (3)] and using our notations, we have

$$E(R) = \max_{p \in \mathbb{P}^n} \min_{Q \in \mathcal{Q}(R)} J(Q, P), \tag{10}$$

where

$$\mathcal{Q}(R) = \{Q : I(p, Q) \leqslant R\}.$$

A useful identity for the average mutual information is

$$I(p, Q) = \min_{q \in \mathcal{P}^n} \bar{J}(Q, q), \tag{11}$$

where

$$\bar{J}(Q, q) := \sum_{k=1}^{m} \sum_{j=1}^{n} p_j Q(k|j) \log[Q(k|j)/q_k].$$

This can be verified by observing that the minimum is achieved for

$$q_k^* = \sum_j p_j Q_{kj}.$$

Using (11), problem (10) can be reformulated as

$$E(R) = \max_{p \in \mathbb{P}^n} \min_{Q \in \mathcal{Q}(R)} \{J(Q, P): \min_{q \in \mathbb{P}^n} \bar{J}(Q, q) \leq R\}. \tag{12}$$

Now, it is an easy exercise to show that any optimization problem of the form

$$\min_x \{f(x): \min_y g(x, y) \leq r\}$$

is equivalent to

$$\min_{x,y} \{f(x): g(x, y) \leq r\};$$

hence, (12) becomes

$$E(R) = \max_p \min_q \min_Q \{J(Q, P): \bar{J}(Q, q) \leq R\}. \tag{13}$$

The inner minimum in (13) is of the form of $e(r)$ in problem (E) and is appropriately denoted by $e(R, q)$. Then, by Theorem 3.1, a dual representation of it is easily shown to be

$$e(R, q) = \max_{\delta \geq 0} \left\{ -\delta R - \log \left\{ \sum_k \sum_j p_j P_{kj}^{1/(i+\delta)} q_k^{\delta/(1+\delta)} \right\}^{1+\delta} \right\}. \tag{14}$$

Substituting the latter representation in (13), we get

$$E(R) = \max_p \min_q \max_{\delta \geq 0} \{g(q, \delta) - \delta R\}, \tag{15}$$

where

$$g(q, \delta) := -\log \left\{ \sum_k \sum_j p_j P_{kj}^{1/(1+\delta)} q_k^{\delta/(1+\delta)} \right\}^{1+\delta}. \tag{16}$$

We shall prove that the min–max appearing in (15) can be reversed. Before that, we need an auxiliary result.

Lemma 3.1. The function $g(q, \delta)$ defined in (16) is

(a) concave in δ, for any $q \in \mathbb{P}^n$,
(b) convex in q, for any $\delta \geq 0$.

Proof. (a) It is well known that the Lagrangian dual function is always concave in the dual variables; hence, (a) follows.

(b) Let $f:\mathbb{R}\to\mathbb{R}$ be a convex decreasing function, and let $g:\mathbb{R}^n\to\mathbb{R}$ be a concave function. Then, it is easy to verify that $h(x)=f(g(x))$ is convex. Take

$$f(t)=-\log t$$

(convex decreasing),

$$g(q)=\sum_k a_k q_k^{\delta/(1+\delta)}, \qquad a_k := \sum_j p_j P_{kj}^{\delta/(1+\delta)} > 0$$

(concave for $\delta \geq 0$). Then, clearly,

$$g(q,\delta)=(1+\delta)f(g(q)),$$

and (b) is proved. □

The min-max theorem related to (15) now follows.

Theorem 3.2. Let $K(q,\delta)=g(q,\delta)-\delta R$. Then, we have

$$\min_q \max_{\delta\geq 0} K(q,\delta) = \max_{\delta\geq 0} \min_q K(q,\delta). \tag{17}$$

Proof. By Lemma 3.1, $K(q,\delta)$ is a convex-concave saddle function for every $q\in\mathbb{P}^n$ and every $\delta\geq 0$. By a result of Rockafeller (Ref. 19), a sufficient condition for the validity of (17) for a general convex-concave saddle function is that

$$\exists \delta_0 \geq 0, \text{ such that } \delta_0 \frac{dK}{d\delta}(q,\delta)\geq 0,\ q\in\mathbb{P}^n,\ \delta>0.$$

This is certainly satisfied if

$$\exists q, \exists \delta>0, \text{ such that } \frac{dK}{d\delta}(q,\delta)<0,$$

i.e.,

$$\exists q, \exists \delta>0, \text{ such that } g'(q,\delta)=\frac{d}{d\delta}g(q,\delta)<R. \tag{18}$$

Since $R>0$, it suffices to prove that

$$\inf_{\delta\geq 0} g'(q,\delta)\leq 0. \tag{19}$$

But $g'(q,\delta)$ is a derivative of a concave function and thus is decreasing; hence,

$$\inf_{\delta\geq 0} g'(q,\delta)=\lim_{\delta\to\infty} g'(q,\delta). \tag{20}$$

Moreover, the gradient inequality for the concave function $g(q, \cdot)$ implies

$$0 = g(q, 0) \leqslant g(q, \delta) - \delta g'(q, \delta),$$

hence

$$g'(q, \delta) \leqslant g(q, \delta)/\delta.$$

Thus, to prove (19), it suffices to show that

$$\lim_{\delta \to \infty} g(q, \delta)/\delta \leqslant 0.$$

Indeed, straightforward computation shows that

$$\lim_{\delta \to \infty} g(q, \delta)/\delta = 0. \qquad \square$$

The last theorem permits us to write $E(R)$ [see, Eq. (15)] as

$$E(R) = \max_{p} \max_{s \geqslant 0} \min_{q} K(q, \delta).$$

However, the next result will show that the inner minimum can be computed, and thus $E(R)$ can be expressed simply as a double maximum problem.

Lemma 3.2. We have

$$\max_{x \in X} \log[\sum x_i^{\alpha/(1+\alpha)} y_i]^{1+\alpha} = \log \sum y_i^{1+\alpha}, \qquad \alpha > 0,$$

where

$$X = \left\{ x \in \mathbb{R}^n : x_k \geqslant 0, \sum_{k=1}^{n} x_k = 1 \right\}.$$

Proof. From the Hölder inequality, we get

$$[\sum x_k^{\alpha/(1+\alpha)} y_k]^{1+\alpha} \leqslant (\sum x_k)^{\alpha} (\sum y_k^{1+\alpha}).$$

Taking logarithms of both expressions and using the fact that $\sum x_k = 1$, we get

$$\sup_{x \in X} \log[\sum x_k^{\alpha/(1+\alpha)} y_k]^{1+\alpha} \leqslant \log \sum y_k^{1+\alpha},$$

and the sup is attained for

$$x_k^* = y_k^{1+\alpha} / \sum y_k^{1+\alpha}. \qquad \square$$

Now, since

$$\min_{q} K(q, \delta) = -\delta R - \max_{q} g(\delta, q),$$

using Lemma 3.2 with

$$x_k := q_k, \qquad y_k := \sum_j p_j P_{kj}^{1/(1+\delta)},$$

a final expression for the reliability rate function $E(R)$ is as follows:

$$E(R) = \max_{p \in \mathbb{R}p^n} \max_{\delta \geq 0} \left\{ -\delta R - \log \sum_k \left[\sum_j p_j P_{kj}^{1/(1+\delta)} \right]^{1+\delta} \right\}. \tag{21}$$

This result coincides with Theorem 18 given in Ref. 14. The second term in (21) is the so-called Gallager function. The dual representation (21) is useful for deriving efficient computational algorithms; see, e.g., Ref. 20.

References

1. KULLBACK, S., and LEIBLER, R. A., *On Information and Sufficiency*, Annals of Mathematical Statistics, Vol. 22, pp. 79–86, 1951.
2. KULLBACK, S., *Information Theory and Statistics*, John Wiley and Sons, New York, New York, 1959.
3. AKAIKE, H., *Information Theory and an Extension of the Maximum Likelihood Principle*, Proceedings of the 2nd International Symposium of Information Theory, Berkeley, California, 1972.
4. CHARNES, A., RAIKE, W. M., and BETTINGER, C. O., *An External and Information Theoretic Characterization of Some Interzonal Transfers*, Socio-Economic Planning Sciences, Vol. 6, pp. 531–537, 1972.
5. CHARNES, A., COOPER, W. W., and LEARNER, D. B., *Constrained Information Theoretic Characterizations in Consumer Purchase Behavior*, Journal of the Operational Research Society, Vol. 29, pp. 833–840, 1978.
6. CHARNES, A., and COOPER, W. W., *An Extremal Principle for Accounting Balance of a Resource-Value Transfer Economy: Existence, Uniqueness, and Computations*, Accademia Nazionale dei Lincei, Series 8, Vol. 56, pp. 556–561, 1974.
7. CHARNES, A., and COOPER, W. W., *Constrained Kullback-Leibler Estimation, Generalized Cobble-Douglas Balance, and Unconstrained Convex Programming*, Accademia Nazionale dei Lincei, Series 8, Vol. 58, pp. 568–576, 1975.
8. SHANNON, C. E., *A Mathematical Theory of Communication*, Bell Systems Technical Journal, Vol. 27, pp. 379–423, 1948 and Vol. 27, pp. 623–656, 1948.
9. BEN-TAL, A., and CHARNES, A., *A Dual Optimization Framework for Some Problems of Information Theory and Statistics*, Problems of Control and Information Theory, Vol. 8, pp. 387–401, 1979.
10. GALLAGER, R. G., *Information Theory and Reliable Communication*, John Wiley and Sons, New York, New York, 1968.
11. JELINEK, F., *Probabilistic Information Theory*, McGraw-Hill. New York, New York, 1968.

12. GALLAGER, R. G., *A Simple Derivation of the Coding Theorem and Some Applications*, IEEE Transactions on Information Theory, Vol. IT-11, pp. 3–18, 1965.
13. CSISZAR, I., and KORNER, J., *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, New York, 1981.
14. BLAHUT, R. E., *On Hypothesis Testing and Information Theory*, IEEE Transactions on Information Theory, Vol. IT-20, pp. 405–417, 1976.
15. ROCKAFELLER, R. T., *Duality and Stability in Extremum Problems Involving Convex Functions*, Pacific Journal of Mathematics, Vol. 21, pp. 167–186, 1976.
16. ROCKAFELLER, R. T., *Conjugate Duality and Optimization*, SIAM Regional Conference Series in Applied Mathematics, Vol. 16, 1974.
17. LAURENT, P. J., *Optimization et Approximation*, Hermann, Paris, Paris, France, 1972.
18. PONSTEIN, J., *Approaches to the Theory of Optimization*, Cambridge University Press, Cambridge, England, 1980.
19. ROCKAFELLER, R. T., *Minimax Theorems and Conjugate Saddle Functions*, Mathematica Scandinava, Vol. 14, pp. 151–173, 1964.
20. ARIMOTO, S., *Computation of Random Coding Exponent Functions*, IEEE Transactions on Information Theory, Vol. IT-22, pp. 665–671, 1976.