Relative Entropy and the Convergence of the Posterior and Empirical Distributions under Incomplete and Conflicting Information

A. Ben-Tal
Faculty of Industrial and Engineeriing and Management
Technion Israel Institute of Technology, and University of Michigan
Haifa, Israel and Ann Arbor, MI 48109

D.E. Brown
Department of Systems Engineering
University of Virginia
Charlottesville, VA 22901

R.L. Smith
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109

Technical Report 88-12
March 1988

Relative Entropy and the Convergence of the Posterior and Empirical Distributions under Incomplete and Conflicting Information

A. Ben-Tal
Faculty of Industrial Engineering and Management
Technion Israel Institute of Technology
Haifa, Israel
and
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109

D.E. Brown*
Department of Systems Engineering
University of Virginia
Charlottesville, VA 22901

R.L. Smith
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109

March 22, 1988

Abstract

It is well known that under complete information the posterior distribution converges to a degenerate distribution concentrated at the true probability distribution as

*The work of this author was partially funded by a grant from The Jet Propulsion Laboratory under Grant #JPL957721.

the sample size grows. Suppose however that the sample possesses average properties not shared in expectation by any probability distribution in the support of the prior. In this case we show that the posterior distribution converges, as sample size grows, to a degenerate distribution p^* closest in relative entropy sense to the set of distributions sharing these average properties. We show moreover that the empirical distribution converges in probability to that empirical distribution v^* closest in relative entropy to the support of the prior. Both p^* and v^* can be obtained computationaly as solutions of certain convex programming problems. Implications for decision making in the presence of conflicting information are explored.

1 Introduction

Consider the problem of estimating a probability distribution when given only partial information on the distribution. When the partial information is based on a random sample drawn from a population following the unknown distribution, a Bayesian construction might provide the mode of the posterior distribution as a best estimate. However, when the partial information consists instead of deterministic constraints on the known distribution, it seems no one best estimate can be singled out from the set of distributions consistent with the information. Jaynes in 1957 introduced a *principle of maximum entropy*, that proposed to choose that probability distribution q of minimum entropy $I(q)$ among those consistent with the constraints. He justified this choice with a *correspondence principle* demonstrating that if the generating distribution were the uniform distribution, then the maximum entropy estimate was the most likely empirical distribution among those consistent with the constraints. This principle has been generalized using the Kullback-Leibler separator $I(q,p)$, between two distributions q and p, to the *principle of minimum relative entropy* (Shore and Johnson [1980]), where this latter principle reduces to the former when p is the uniform distribution. In general, there has been enormous interest in this interplay between statistics and information theory within the communities of statistics (Sanov [1961], Vasicek [1980], and Bahadur [1971]), information theory (Van Campenhout and Cover [1981], Jaynes [1956], Johnson [1979], and Sampson and Smith [1984]), and operations research (Cozzolino and Zahner [1973], Sampson and Smith [1982], Thomas [1979], and Wilson [1970]).

In order to make sense of the correspondence principle, some kind of superpopulation model seems inescapable. That is, the maximum entropy distribution estimates the distribution for a population generated through a random sample from a superpopulation following a uniform distribution. On the other hand, the Bayesian construction, which chooses the mode of the posterior would then be providing an estimate for the distribution of the superpopulation based on knowledge about the population distribution, the latter regarded as a sample drawn from the former. We intend to pursue this dual interpretation within this paper.

Specifically, we suppose that certain constraints are imposed on the population (or empirical) distribution V^n for population size n as well as on the superpopulation distribution P. The former is represented by a subset Λ of distributions known to contain V^n while the latter is represented by the support Ω of the prior distribution $\mathcal{P}$ over P. To avoid the trivial case we suppose that Λ and Ω do not intersect. In section 2, the formal probability model is presented. In section 3 we collect and refine some Large Deviation results, and in section 4 we prove results concerning uniqueness and continuity of the relative entropy minimizers. These results are in preparation for section 5 which contains the main convergence results; in Theorem 5.1 and Corollary 5.2, we show that the posterior distribution $\mathcal{P}^n$, based on a sample of size n, converges (as n goes to infinity) to a degenerate distribution concentrated at the distribution p^* in Ω that minimizes the relative entropy between Λ and Ω. In Theorem 5.4 and Corollary 5.5, under mild regularity conditions, we extend the correspondence principle within this Bayesian construction. We show that not only the mode but the conditional distribution of V^n converges (as n goes to infinity) to a degenerate distribution concentrated at the distribution ν^* in Λ that minimizes the relative entropy between points of Λ and Ω. Both ν^* and p^* in Ω are independent of the prior $\mathcal{P}$ and only depend on the geometry of Λ and Ω. Both p^* and ν^* can be computed as optimal solutions of certain convex programming problems. In section 6, we discuss the implications of these results to decision making under incomplete and conflicting information.

2 The Probability Model

Let the prior distribution $\mathcal{P}$ be an absolutely continuous probability measure with support $\Omega \subset S$ where $S = \{q | \sum_{i=0}^{m} q_i = 1, q_i > 0\}$ is the relative interior of the m-dimensional standard simplex in R^{m+1}. The true distribution P is a random probability mass function $P \in \Omega$ with distribution $\mathcal{P}$. Let x be a finite discrete random variable that takes on the value $a_i \in \Re$ with probability P_i for $i = 0, 1, \ldots, m$.

Suppose P is determined and *fixed* in accordance with the prior distribution $\mathcal{P}$ and X is then repeatedly observed within an experiment of n trials to yield a sample $X_1, X_2, \ldots, X_n$. That is, $X_1, X_2, \ldots, X_n$ are identically distributed with p.m.f. P and conditionally independent when given P. Let N_i be the random variable equal to the number of times outcome a_i is observed in the n trials for $i = 0, 1, 2, \ldots, m$. The *empirical distribution* (based on n trials) V^n is the random vector in R^{m+1} whose ith component represents the relative frequency of occurrence of outcome a_i in n trials, i.e.,

$$V^n = (\frac{N_0}{n}, \frac{N_1}{n}, \ldots, \frac{N_m}{n}).$$

Let Λ be a Borel measurable subset of S with $\Pr(V^n \in \Lambda) > 0$ for n sufficiently large. For Borel sets $B \subset S$, the *posterior distribution* $\mathcal{P}^n$ of P, when given the information that

$V^n \in \Lambda$, is

$$\mathcal{P}^n(B) = \Pr(P \in B | V^n \in \Lambda)$$

and the *sample distribution* $\mathcal{V}^n$ of V^n when given the information that $V^n \in \Lambda$, is

$$\mathcal{V}^n(B) = \Pr(V^n \in B | V^n \in \Lambda).$$

We consider in the next section conditions under which the posterior and sample distributions, $\mathcal{P}^n$ and $\mathcal{V}^n$ converge as the sample size n increases.

3 Preliminary Large-Deviation Results

In this section we derive a Large-Deviation type result (Theorem 3.3) on the conditional probability of the empirical distribution.

We begin by introducing the *relative entropy* $I(v,p)$ between two distribution v and p in S, which is defined by

$$I(v,p) = \sum_{i=0}^{m} v_i \log (v_i/p_i)$$

where *log* is here and elswhere taken to the base *e*. $I(v,p)$ is also called the *Kullback-Leibler separator* (Kullback [1959]) and the *minimum discrimination information* (e.g. Brockett, Charnes and Cooper [1980]). Since $I(v,p)$ is nonnegative, and is equal to zero if and only if $v = p$, it is also referred to as the *directed distance* from p to v. Although I is not a metric (it is not symmetric, nor does it satisfy the triangle inequality) we shall occasionally adopt this distance terminology to draw geometric analogies for the results to follow.

For a given nonempty subset $\Lambda \subset S$, and a probability vector $p \in S$, we denote by $I(\Lambda, p)$ the relative entropy ("distance") between p and Λ, defined by

$$I(\Lambda, p) = \inf_{v \in \Lambda} I(v,p).$$

The main properties of $I(v,p)$ and $I(\Lambda,p)$, needed for our purposes, are summarized in Appendix I. The main result in this section (Theorem 3.3) is preceded with a result on uniform convergence of $I(\Lambda^n, p)$(Lemma 3.1) and a well known result on bounding the conditional probability $\Pr\{V^n \in \Lambda | P = p\}$(Lemma 3.2).

Since the empirical distribution V^n when multiplied by n yields an integer vector, only a discrete lattice of points within the set Λ are possible values for V^n. That is, the set Λ^n of possible values of V^n is given by

$$\Lambda^n = \Lambda \cap S^n$$

where $S^n = \{q \in S | nq_i$ is a nonnegative integer for $i = 0, 1, \ldots, m\}$. Since Λ^n becomes dense in Λ as n goes to infinity (for Λ a closed body,) it is not perhaps surprising that the directed distance from p to Λ^n converges to the directed distance from p to Λ. Lemma 1 below extends this result in Bahadur [1971] by showing that for closed Ω this convergence is *uniform* in p.

Lemma 3.1 *Let Ω and Λ be subsets of S with Ω closed and Λ a closed body (i.e., Λ is the closure of its interior Λ°.) Then*

$$I(\Lambda^n, p) \longrightarrow I(\Lambda, p)$$

uniformly *over $p \in \Omega$ as $n \to \infty$.*

Proof: We first show that $D(\Lambda^n, \Lambda) \to 0$ as $n \to \infty$ where D is the *Hausdorff metric* defined for compact sets F and D by $D(F, G) = \max(h(F, G),\ h(G, F))$ with $h(F, G) = \max\{d(x, G),\ x \in F\}$ and $d(x, y)$ being the Euclidian distance between x and y. Now $D(\Lambda^n, \Lambda) \to 0$ as $n \to \infty$ iff (1) $\limsup_{n\to\infty} \Lambda^n \subseteq \Lambda$ and (2) $\Lambda \subseteq \liminf_{n\to\infty} \Lambda^n$. Here, for a sequence of compact sets $\{F_n\}$, $\limsup_{n\to\infty} F_n$ (resp. $\liminf_{n\to\infty} F_n$) is the set of all points x such that every open neighborhood of x is frequently (resp. eventually) intersected by the F_n. (Hausdorff [1957]).

Now, (1) follows immediately from the fact that $\Lambda^n \subseteq \Lambda$ for all n and Λ is closed. As for (2), let $\nu \in \Lambda$ and let $N(\nu)$ be an open neighborhood of ν. Now $N(\nu) \cap \Lambda^0 \neq \emptyset$ since Λ is the closure of Λ^0 and therefore $N(\nu) \cap \Lambda^0$ being the intersection of two open sets is open. It follows that there exists an $\epsilon > 0$ and $x \in N(\nu) \cap \Lambda^0$ such that the open ball $B_\epsilon(x)$ of radius ϵ around x satisfies $B_\epsilon(x) \subseteq N(\nu) \cap \Lambda^0$. Let x^n be the Euclidian closest point in S^n to x. Since $x \in S, d(x^n, x) \to 0$ as $n \to \infty$ and hence there is an integer N such that $x^n \in B_\epsilon(x)$ for all $n \geq N$. But then $x^n \in \Lambda^n$ since $B_\epsilon(x) \subseteq \Lambda^0 \subseteq \Lambda$ for all $n \geq N$ for some N. Also $x^n \in N(\nu)$ for all $n \geq N$ since $B_\epsilon(x) \subseteq N(\nu)$. Hence $\Lambda^n \cap N(\nu) \neq \emptyset$ for all $n \geq N$ and $\Lambda \subseteq \liminf_{n\to\infty} \Lambda^n$. Therefore $D(\Lambda^n, \Lambda) \to 0$ as $n \to \infty$.

Choose $\epsilon > 0$. Since $\Lambda^n \subset \Lambda$ for all $n, I(\Lambda^n, p) \geq I(\Lambda, p) > I(\Lambda, p) - \epsilon$ for all $p \in \Omega$ and for all n. It remains to show that $I(\Lambda^n, p) < I(\Lambda, p) + \epsilon$ for all $p \in \Omega$ for all $n \geq N_\epsilon$ for some N_ϵ. Let $\nu(p) \in \operatorname{Arg\,min}\{I(\nu, p), \nu \in \Lambda\} \neq \emptyset$ since $I(\nu, p)$ is a continuous function over the compact set Λ. Since $D(\Lambda^n, \Lambda) \to 0$ as $n \to \infty$, for all $\delta > 0$, there is a N'_δ such that $D(\Lambda^n, \Lambda) < \delta$ for all $n \geq N'_\delta$. In particular, for all $\nu \in \Lambda$ there are $q^n(\nu) \in \Lambda^n$ for all $n \geq N'_\delta$ such that $d(q^n(\nu), \nu) < \delta$.

Now since $I(\nu, p)$ is continuous on $\Lambda \times \Omega$ (see Appendix I(d)), which is compact, $I(\nu, p)$ is uniformly continuous over $\Lambda \times \Omega$. In particular, for all $\epsilon > 0$ there is a δ_ϵ independent of $p \in \Omega$ such that if $d(\nu, \nu') < \delta_\epsilon$ for $\nu, \nu' \in \Lambda$ then $|I(\nu, p) - I(\nu', p)| < \epsilon$. Let $N_\epsilon = N'_{\delta_\epsilon}$ so that $d(q^n(\nu(p)), \nu(p)) < \delta_\epsilon$ for all $n \geq N_\epsilon$. Then $I(q^n(\nu(p), p) - I(\nu(p), p) < \epsilon$ for all $p \in \Omega$ for all $n \geq N_\epsilon$. Hence $I(\Lambda^n, p) - I(\Lambda, p) \leq I(q^n(\nu(p)), p) - I(\nu(p), p) < \epsilon$ for all $p \in \Omega$ for all $n \geq N_\epsilon$. Therefore $I(\Lambda^n, p) < I(\Lambda, p) + \epsilon$ and the result follows. ∎

The following well known lemma provides bounds on the probability of finding the empirical distribution in $\Lambda \subset S$, when given P, in terms of the directed distance from P to Λ^n.

Lemma 3.2 *Let Λ be a subset of S and $p \in S \subset R^{m+1}$. Then there exists a positive constant $\gamma(m)$, depending only on m, such that for all n*

$$\exp\{-n[I(\Lambda^n,p)+(\frac{m-1}{2})\frac{\log\ n}{n}-\frac{\log\gamma(m)}{n}]\} \leq \Pr(V^n \in \Lambda|P=p)$$

$$\leq \exp\{-n[I(\Lambda^n,p)-m\frac{\log(n+1)}{n}]\}.$$

Proof: See Bahadur [1971,p.18]. ∎

Combining Lemmas 3.1 and 3.2, we get the following important large-deviation result.

Theorem 3.3 *Let Ω and Λ be subsets of S with Ω closed and Λ a closed body. Then*

$$\frac{1}{n}\ln\Pr(V^n \in \Lambda|P=p) \longrightarrow -I(\Lambda,p)$$

uniformly over $p \in \Omega$ as $n \to \infty$.

Proof: From Lemma 3.2,

$$-(I(\Lambda^n,p)+(\frac{m-1}{2})\frac{\log n}{n}-\frac{\log\gamma(m)}{n}) \leq \frac{1}{n}\log\Pr(V^n \in \Lambda|P=p) \leq -(I(\Lambda^n,p)-m\frac{\log(n+1)}{n})$$

for all n, and hence

$$I(\Lambda,p)-I(\Lambda^n,p)-(\tfrac{m-1}{2})\tfrac{\log n}{n}+\tfrac{\log\gamma(m)}{n} \leq \tfrac{1}{n}\ln\Pr(V^n \in \Lambda|P=p)+I(\Lambda,p)$$
$$\leq I(\Lambda,p)-I(\Lambda^n,p)+m\tfrac{\log(n+1)}{n}$$

for all n. From Lemma 3.1, $I(\Lambda^n,p) \to I(\Lambda,p)$ uniformly over $p \in \Omega$ as $n \to \infty$. Let $\epsilon > 0$ and choose N_1, N_2, N_3 large enough so that

$$|I(\Lambda,p)-I(\Lambda^n,p)| < \frac{\epsilon}{2}$$

for all $p \in \Omega$ and for all $n \geq N_1$,

$$-(\frac{m-1}{2})\frac{\log n}{n}+\frac{\log\gamma(m)}{n} > -\frac{\epsilon}{2}$$

for all $n \geq N_2$, and

$$m\frac{\log(n+1)}{n} < \frac{\epsilon}{2}$$

for all $n \geq N_3$. Let $N = \max(N_1, N_2, N_3)$. Then

$$-\epsilon < \frac{1}{n}\log \Pr(V^n \in \Lambda | P = p) + I(\Lambda, p) < \epsilon$$

for all $p \in \Omega$ and all $n \geq N$. ∎

Theorem 3.3 states that for large n,

$$P(V^n \in \Lambda | P = p) \approx \exp\{-nI(\Lambda, p)\}.$$

Hence for $p \notin \Lambda$, the probability of finding the empirical distribution in a set Λ decays exponentially in its directed distance from p. Thus the likelihood of finding the empirical distribution in the closer to p of two sets is overwhelmingly greater then that of finding it in the more distant one. By demonstrating that the convergence is uniform over closed Ω, Theorem 3.3 is an extension of large deviation theory that began with the classic work of Sanov [1961].

4 Uniqueness and Continuity of the Relative Entropy Minimizers

The convergence results in the next section make use of uniqueness and continuity properties of the optimal solution sets

$$\{p^* | p^* = \arg\min_{v \in \Lambda} I(v,p)\}, \{v^* | v^* = \arg\min_{p \in \Omega} I(v,p)\} \text{ and}$$
$$\{(v^*, p^*) | I(v^*, p^*) = \min_{\Lambda \times \Omega} I(v,p)\}.$$

In this section we study these properties. The first lemma gives necessary conditions for nonuniqueness of the solution of

$$\min_{\Lambda \times \Omega} I(v,p).$$

Lemma 4.1 *Let Λ, Ω be convex subsets of S. Consider the problem (P)*

$$(P) \qquad \min_{v \in \Lambda, p \in \Omega} \{I(v,p) = \sum_{i=0}^{m} v_i \log(\frac{v_i}{p_i}\}.$$

If $(\bar{v}, \bar{p})$ and (v^*, p^*) are both solutions of (P), then

$$\frac{\bar{v}_i}{\bar{p}_i} = \frac{v_i^*}{p_i^*}, \text{ all } i = 0, 1, \ldots, m. \tag{1}$$

Proof: Let $I^* = I(v^*, p^*) = I(\bar{v}, \bar{p}) = \min I(v, p)$. Since I is a convex function on $R^n_+ \times R^n_+$ (see Appendix I (a)) and the feasible set of $(P), \Lambda \times \Omega$, is convex, it follows that the set of optimal solutions of (P) is convex. In particular then $\forall 0 \leq \alpha \leq 1, (v^\alpha, p^\alpha) = (\alpha v^* + (1-\alpha)\bar{v}, \alpha p^* + (1-\alpha)\bar{p})$ is an optimal solution, i.e.,

$$I(\alpha) \equiv I(v^\alpha, p^\alpha) = I^*, \quad \text{all } 0 \leq \alpha \leq 1.$$

The latter equation implies

$$I'(0) = 0 \tag{2}$$

i.e.,

$$\begin{aligned} I'(0) &= \textstyle\sum_{i=0}^m \frac{d}{d\alpha}\{(\alpha v_i^* + (1-\alpha)\bar{v}_i)\log(\frac{\alpha v_i^* + (1-\alpha)\bar{v}_i}{\alpha p_i^* + (1-\alpha)\bar{p}_i})\}|_{\alpha=0} \\ &= \textstyle\sum_{i=0}^m \{(v_i^* - \bar{v}_i)\log(\bar{v}_i/\bar{p}_i) - v_i^* - \bar{v}_i - (p_i^* - \bar{p}_i)(\bar{v}_i/\bar{p}_i)\} = 0. \end{aligned}$$

Hence, using $\sum \bar{v}_i = \sum v_i^* = 1$, and (2),

$$\textstyle\sum_{i=0}^m \{v_i^* \log(\bar{v}_i/\bar{p}_i) - (\bar{v}_i p_i^*/\bar{p}_i\} - I^* + 1 = 0. \tag{3}$$

Consider the one-variable problem

$$\max_{v_i > 0} \{v_i^* \log(v_i/\bar{p}_i) - v_i p_i^*/\bar{p}_i\}.$$

The *unique* optimal solution of this strictly concave problem is

$$\hat{v}_i = \frac{v_i^* \bar{p}_i}{p_i^*}$$

and the corresponding optimal value is

$$v_i^* \log(v_i^*/p_i^*) - v_i^*.$$

Hence

$$\begin{aligned} I^* - 1 = \sum_{i=0}^m \{v_i^* \log(v_i^*/p_i^*) - v_i^*\} &= \max_{v_i > 0} \{\textstyle\sum v_i^* \log(v_i/\bar{p}_i) - v_i(p_i^*/\bar{p}_i)\} \\ > \sum_{i=0}^m \{v_i^* \log(\bar{v}_i/\bar{p}_i) - (\bar{v}_i p_i^*/\bar{p}_i)\}, &\text{if } \bar{v}_j \neq \hat{v}_j, \text{ for some } j, \\ = I^* - 1 &\text{ by (3)}, \end{aligned}$$

which is a contradiction. Hence $\bar{v}_i = \hat{v}_i = v_i^* \bar{p}_i / p_i^* \ \forall i = 0, 1, \ldots, m$ proving (1). ∎

The *necessary* conditions for *nonuniqueness* can generate sufficient conditions for uniqueness. This is done in the next three results. The first result is for the two-dimensional case $(m = 1)$.

Corollary 4.2 *Let Λ, Ω be convex subsets of $S \subset R^2$, with $\Lambda \cap \Omega = \emptyset$. Then problem (P) has a unique solution.*

Proof: Suppose there are two distinct solutions $(\lambda^*, p^*) \neq (\bar{\lambda}, \bar{p})$. Since $\Lambda \cap \Omega = \emptyset, \lambda^* \neq p^*$ i.e.,

$$\lambda_0^* \neq p_0^*. \tag{4}$$

By Lemma 4.1, we must have

$$\left.\begin{array}{ll} \lambda_0^*/p_0^* = \bar{\lambda}_0^*/\bar{p}_0, & \lambda_1^*/p_1^* = \bar{\lambda}_1/\bar{p}_1, \\ \bar{\lambda}_0 + \bar{\lambda}_1 = 1, & \bar{p}_0 + \bar{p}_1 = 1. \end{array}\right\} \tag{5}$$

Consider the system of linear equations

$$\begin{bmatrix} p_0^* & -\lambda_0^* & 0 & 0 \\ 0 & 0 & p_1^* & -\lambda_1^* \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \lambda_0 \\ p_0 \\ \lambda_1 \\ p_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}. \tag{6}$$

Notice that $(\lambda_0^*, p_0^*, \lambda_1^*, p_1^*)$ is a solution, as well as, by (5), $(\bar{\lambda}_0, \bar{p}_0, \bar{\lambda}_1, \bar{p}_1)$. But this is a contradiction since (4) implies that the coefficient matrix in (6) is nonsingular. ∎

An important special case of the problem (P) is where Ω(or Λ) consists of a single probability mass function. In this case uniqueness is also obtained.

Corollary 4.3 *If one of the sets $\Lambda, \Omega \subset S \subset R^{m+1}$ is a singleton, and the other is convex, and $\Lambda \cap \Omega = \emptyset$, then problem (P) has a unique solution.*

Proof: Let $\Omega = \{p^0\}$. If problem (P) has two solutions they are of the form (v^*, p^0) and $(\bar{v}, p^0)$. and by Lemma 4.1 they must satisfy

$$\frac{v_i^*}{p_i^0} = \frac{\bar{v}_i}{p_i^0} \qquad \forall i = 0, 1, \ldots, m.$$

Hence $v^* = \bar{v}$ and the two solutions coincide. ∎

The third corollary gives a reasonable general sufficient condition for uniqueness. For its formulation we need the notion of a *strictly convex set.* Thus a convex set C is strictly convex if int $C \neq \emptyset$ and $\forall x, y \in C$ and $0 < \lambda < 1$ it follows that $\lambda x + (1 - \lambda)y \in$ int C.

A typical example of such sets are level sets

$$C = \{x \in R^n | f(x) \leq \gamma\}$$

where f is a strictly convex function and $\gamma > \inf(f)$.

Corollary 4.4 *If one of the sets Λ, Ω is convex and the other is strictly convex and $\Lambda \cap \Omega = \emptyset$, then problem (P) has a unique solution.*

Proof: Suppose (v^*, p^*) and $(\bar{v}, \bar{p})$ are distinct solutions of (P). Then by Lemma 4.1 we must have

$$v^* \neq \bar{v} \text{ and } p^* \neq \bar{p}.$$

Suppose that Ω is strictly convex. Let

$$\hat{v} = \frac{1}{2}(v^* + \bar{v}), \quad \hat{p} = \frac{1}{2}(p^* + \bar{p})$$

and note that $(\hat{v}, \hat{p}) \in \Lambda \times \Omega$ is also an optimal solution of (P), i.e.,

$$I(\hat{v}, \hat{p}) = \min_{v \in \Lambda, p \in \Omega} I(v, p) = I^*.$$

Since Ω is strictly convex, $\hat{p} \in \text{int } \Omega$. Also $\Lambda \cap \Omega = \emptyset$ implies that $\hat{v} \notin \Omega$. Hence there exists $0 < \gamma < 1$ such that the point $p_\gamma = \gamma \hat{v} + (1 - \gamma)\hat{p} \in \Omega$. Therefore $(\hat{v}, p_\gamma) \in \Lambda \times \Omega$ and so

$$I^* \leq I(\hat{v}, p_\gamma) < \gamma I(\hat{v}, \hat{v}) + (1 - \gamma) I(\hat{v}, \hat{p}) = \gamma \cdot 0 + (1 - \gamma) I^* < I^*,$$

a contradiction. Thus no two distinct solutions exist. The same proof goes through if Λ is assumed to be strictly convex. ∎

If none of the three sufficient conditions in Corollaries 4.1 through 4.3 hold, problem (P) may indeed have more than one optimal solution. This is illustrated by the following.

Example Consider the special case of problem (P):

$$\min_{v \in \Lambda, p \in \Omega} \sum_{i=0}^{2} v_i \log(v_i / p_i)$$

where $\Lambda = \{v \in R^3 | v_0 + 2v_1 + v_2 \leq 4/3, v \geq 0, \sum v_i = 1\}$ and $\Omega = \{p \in R^3 | p_0 + 2p_1 + p_2 \geq 3/2, p \geq 0, \sum p_i = 1\}$.

Hence Λ and Ω are convex subsets of R^3, $\Lambda \cap \Omega = \emptyset$, but neither Λ nor Ω are strictly convex (or a singleton).

It is easily verified from the Karush-Kuhn-Tucker conditions that

$$v^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), p^* = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$$

and

$$\bar{v} = (\frac{1}{6}, \frac{1}{3}, \frac{1}{2}), \bar{p} = (\frac{1}{8}, \frac{1}{2}, \frac{3}{8})$$

are both optimal solutions.
Note that the necessary condition (1) is satisfied here:

$$\frac{v_i^*}{p_i^*} = \frac{\bar{v}_i}{\bar{p}_i},\ i = 0, 1, 2. \quad \blacksquare$$

The discussion of the uniqueness has been limited so far to the case where both Λ and Ω are convex sets. Without this assumption, the uniqueness hypothesis which is needed for the convergence results in Section 5 is as follows:

A1 $\forall p \in \Omega$, there is a unique solution $v(p)$ where

$$v(p) = \arg\min_{v \in \Lambda} I(v, p)$$

and

A2 There is a unique solution p^* where

$$p^* = \arg\min_{p \in \Omega} I(v(p), p).$$

However, under the convexity assumption, the uniqueness condition $A1 \cap A2$ is equivalent to the uniqueness of the optimal solution to problem (P).

Lemma 4.5 *Let Λ and Ω be convex, then $A1 \cap A2$ holds if and only if condition B holds:*

B The problem $min_{v \in \Lambda, p \in \Omega} I(v, p)$ has a unique solution.

Proof: $(B \Longrightarrow A1 \cap A2)$ Suppose A1 is false, i.e., there exists $\bar{p} \in \Omega$ and $v_1, v_2 \in \Omega, v_1 \neq v_2$ such that $v_i \in \text{Arg}\min_{v \in \Lambda} I(v, \bar{p})$. But this is impossible since $I(\cdot, \bar{p})$ is strictly convex, and hence possesses a unique solution on the convex set Λ. Suppose then that $A1$ holds but not $A2$, i.e. for some $p_1^*, p_2^*, \in \Omega, p_1^* \neq p_2^*, p_i^* \in \text{Arg}\min_{p \in \Omega} I(v(p), p)$. Let $v_i^* = v(p_i^*)$. Then

$$I(v_1^*, p_1^*) = I(v_2^*, p_2^*) = \min I(v, p). \tag{7}$$

For otherwise there exists $(\bar{v}, \bar{p}) \in \Lambda \times \Omega$ such that

$$I(\bar{v}, \bar{p}) < I(v(p_i^*), p_i^*). \tag{8}$$

Now, $I(v(p_i^*), p_i^*) \leq I(v(\bar{p}), \bar{p})$ since $p_i^* = \arg\min I(v(p), p) \leq I(\bar{v}, \bar{p}) < I(v(p_i^*), p_i^*)$ by (8), a contradiction. Hence (7) holds, but this is in turn a contradiction to B.

$(A1 \cap A2 \Longrightarrow B)$ Suppose B is false and let (v_1^*, p_2^*) and (v_2^*, p_2^*) be both solutions of $\min_{v\in\Lambda, p\in\Omega} I(v,p)$ with

$$(v_1^*, p_1^*) \neq (v_2^*, p_2^*). \tag{9}$$

By Lemma 4.1, $\frac{(v_1^*)_i}{(p_1^*)_i} = \frac{(v_2^*)_i}{(p_2^*)_i}$ $\forall i$. Hence, (9) implies

$$v_1^* \neq v_1^*, p_1^* \neq p_2^*.$$

Now, by $A1, v_i^* = v(p_i^*)$ and so both p_1^* and p_2^* are solutions of $\min I(v(p),p)$, with $p_1^* \neq p_2^*$, but this contradicts $A2$. ∎

Whenever assumption $A1$ holds, the mapping

$$p \rightarrow \arg\min_{v\in\Lambda} I(v,p)$$

is single-valued and defines a function $v(\cdot)|\Omega \rightarrow \Lambda$. The last lemma is a general result, directly applicable to the question of the continuity of the function $v(p)$.

Lemma 4.6 *Let X be a compact subset of R^n and $Y \subset R^m$. Consider the function $f|R^n \times R^m \rightarrow R$ and assume that f is continuous on $X \times Y$ and that $\forall y \in Y$, there exists a unique solution*

$$x(y) = \arg\min_{y\in Y} f(x,y).$$

Then, $x(\cdot)$ is a continuous function on Y.

Proof: Let $y_n \rightarrow y$ and suppose, for some open neighborhood $N(x(y))$ of $x(y)$, that $x(y_{n_k}) \notin N(x(y))$ for all k. Passing to a subsequence if necessary, we have $x(y_{n_k}) \rightarrow x'$ for some $x' \in X$, by compactness of X. Also $I(x(y),y) < I(x',y)$ since $x(y)$ is the unique minimizer for y. Let $\epsilon = f(x',y) - f(x(y),y) > 0$. Let k_1 be such that $|f(x(y),y_{n_k}) - f(x(y),y)| < \epsilon/2$ for all $k \geq k_1$, which is possible since $f(x,\cdot)$ is continuous on Y. Let k_2 be such that $|f(x',y) - f(x(y_{n_k}),y_{n_k})| < \epsilon/2$ for all $k \geq k_2$ which is possible since f is continuous on $X \times Y$. We have $f(x(y),y_{n_k}) - f(x(y_{n_k}),y_{n_k}) \leq |f(x(y),y_{n_k}) - f(x(y),y)| + f(x(y),y) - f(x',y) + |f(x',y) - f(x(y_{n_k}),y_{n_k})| < \epsilon/2 - \epsilon + \epsilon/2 = 0$ for $k > \max(k_1,k_2)$. Hence for such k, we have $f(x(y),y_{n_k}) < f(x(y_{n_k}),y_{n_k})$ which is a contradiction to $x(y_{n_k})$ being the unique minimizer of $f(\cdot,y_{n_k})$ on X. ∎

5 Convergence of the Posterior and Empirical Distributions

Roughly speaking, if we are given that the empirical distribution is in Λ, Theorem 3.1 tells us that we would find the highest likelihood for it being in a neighborhood of the point

$v(p)$ closest (in relative entropy distance) to $p \in \Omega$. Since this likelihood increases as p gets closer to $v(p)$, we should expect that P has the highest probability of being near the point p^* closest to Λ. This is indeed the case as expressed in the following.

Theorem 5.1 (Posterior Convergence) *Let the prior $\mathcal{P}$ be an absolutely continuous probability measure with support $\Omega \subset S$ where Ω is a closed set. Let $\Lambda \subset S$ be a closed body with $\Lambda \cap \Omega = \emptyset$. Suppose $\nu(p) = \arg\min_{\nu \in \Lambda} I(\nu, p)$ is unique for all $p \in S$ and $p^* = \arg\min_{p \in \Omega} I(\nu(p), p)$ is also unique. Then*

$$\mathcal{P}^n \stackrel{d}{\Longrightarrow} p^* as\ n \to \infty$$

in the sense that $\mathcal{P}^n(N(p^*)) \to 1$ as $n \to \infty$ for all open neighborhoods $N(p^*)$ of p^*.

Proof: We have

$$\begin{aligned}
\Pr(V^n \in \Lambda | P \notin N(p^*)) &\leq \sup_{p \in \Omega - N(p^*)} \Pr(V^n \in \Lambda | P = p) \\
&\leq \sup_{p \in \Omega - N(p^*)} \exp\{-n(I(\Lambda^n, p) - m\frac{\log(n+1)}{n})\} \text{ by Lemma 3.2,} \\
&\leq \sup_{p \in \Omega - N(p^*)} \exp(-n(I(\Lambda, p) - m\frac{\log(n+1)}{n}) \text{ since } \Lambda^n \subset \Lambda \text{ for all } n. \\
&= \exp\{-n(I(\nu(\underline{p}), \underline{p}) - m\frac{\log(n+1)}{n})\}
\end{aligned}$$

where

$$\underline{p} \in \text{Arg} \min_{p \in \Omega - N(p^*)} I(\nu(p), p)$$

which exists since $I(\nu(p), p)$ being a composition of continuous functions (see Lemma 4.6) is continuous over the compact set $\Omega - N(p^*)$.
Also,

$$\begin{aligned}
\Pr(V^n \in \Lambda) &= \int_\Omega \Pr(V^n \in \Lambda | P = p) d\mathcal{P}(p) \\
&\geq \int_\Omega \exp\{-n(I(\Lambda^n, p) + (\frac{m-1}{2})\frac{\log n}{n} - \frac{\log \gamma(m)}{n})\} d\mathcal{P}(p) \text{ by Lemma 3.2.} \\
&\geq \int_{\Omega_\epsilon(p^*)} \exp\{-n(I(\Lambda^n, p) + (\frac{m-1}{2})\frac{\log n}{n} - \frac{\log \gamma(m)}{n})\} d\mathcal{P}(p)
\end{aligned}$$

since $\Omega_\epsilon(p^*) \subset \Omega$, where

$$\Omega_\epsilon(p^*) = \Omega \cap \{q \in S | |q_i - p_i^*| < \epsilon \text{ for } i = 0, 1, 2, \ldots, m\}.$$

By Lemma 3.1, $I(\Lambda^n, p) \to I(\Lambda, p)$ uniformly over p in the closure of $\Omega_\epsilon(p^*)$, so that for all $\delta > 0$ there exists an N_δ independent of $p \in \Omega_\epsilon(p^*)$ such that

$$I(\Lambda^n, p) \leq I(\Lambda, p) + \delta$$

for all $n \geq N_\delta$ and all $p \in \Omega_\epsilon(p^*)$.
Therefore, by the above inequality, we get

$$\begin{aligned}
\Pr(V^n \in \Lambda) &\geq \int_{\Omega_\epsilon(p^*)} \exp\{-n(I(\Lambda, p) + \delta + (\frac{m-1}{2})\frac{\log n}{n} - \frac{\log \gamma(m)}{n})\} d\mathcal{P}(p) \\
&\geq \int_{\Omega_\epsilon(p^*)} \exp\{-n(I(\nu(\bar{p}(\epsilon)), \bar{p}(\epsilon)) + \delta + \\
&\quad (\frac{m-1}{2})\frac{\log n}{n} - \frac{log\gamma(m)}{n})\} d\mathcal{P}(p) \qquad (10)
\end{aligned}$$

for all $n > N_\delta$, for all $\delta > 0$, where

$$\bar{p}(\epsilon) \in \text{Arg} \min_{p \in \bar{\Omega}_\epsilon(p^*)} I(v(p), p)$$

and $\bar{\Omega}_\epsilon(p^*)$ is the closure of $\Omega_\epsilon(p^*)$. Hence

$$\Pr(V^n \in \Lambda) \geq \exp\{-n(I(\nu(\bar{p}(\epsilon)), \bar{p}(\epsilon)) + \delta + (\frac{m-1}{2})\frac{\log n}{n} - \frac{\log \gamma(m)}{n})\} \mathcal{P}(\Omega_\epsilon(p^*)) > 0$$

for all $n \geq N_\delta$, for all $\epsilon, \delta > 0$ since $\mathcal{P}$ is absolutely continuous over Ω.
Hence, using the two inequalities proved above,

$$\begin{aligned}
\Pr(P \notin N(p^*) | V^n \in \Lambda) &= \frac{\Pr(V^n \in \Lambda | P \notin N(p^*)) \Pr(P \notin N(p^*))}{\Pr(V^n \in \Lambda)} \\
&\leq \frac{\exp\{-n(I(\nu(\underline{p}), \underline{p}) - m\frac{\log(n+1)}{n})\}(1 - \mathcal{P}(N(p^*))}{\exp\{-n(I(\nu(\bar{p}(\epsilon)), \bar{p}(\epsilon)) + \delta + (\frac{m-1}{2})(\frac{\log n}{n}) - \frac{\log \gamma(m)}{n})\} \mathcal{P}(\Omega_\epsilon(p^*))} \\
&= \exp\{-n(I(\nu(\underline{p}), \underline{p}) - I(\nu(\bar{p}(\epsilon)), \bar{p}(\epsilon)) - \delta - (\frac{m-1}{2})(\frac{\log n}{n}) \\
&\quad + \frac{\log \gamma(m)}{n} - m\frac{\log(n+1)}{n})\} \cdot \frac{1 - \mathcal{P}(N(p^*))}{\mathcal{P}(\Omega_\epsilon(p^*))}
\end{aligned}$$

for all $n \geq N_\epsilon$ for all $\delta, \epsilon > 0$.
Now choose $\epsilon > 0$ sufficiently small so that $I(\nu(\underline{p}), \underline{p}) > I(\nu(\bar{p}(\epsilon)), \bar{p}(\epsilon))$ which is possible since $I(\nu(p), p)$ is a continuous function of $p \in S$ and $p^* = \arg\min_{p \in \Omega} I(\nu(p), p)$ is unique.

Let $\delta = (I(\nu(\underline{p}), \underline{p}) - I(\nu(\bar{p}(\epsilon)), \bar{p}(\epsilon)))/2 > 0$.
Then

$$\Pr(P \notin N(p^*) | V^n \in \Lambda) \leq \exp\{-n(\delta + \frac{\log \gamma(m)}{n} - \frac{(m-1)}{2}(\frac{\log n}{n}) - m\frac{\log(n+1)}{n})\}\frac{1 - \mathcal{P}(N(p^*))}{\mathcal{P}(\Omega_\epsilon(p^*))}$$

for all $n \geq N_\delta$.
Hence $\limsup_{n\to\infty} \Pr(P \notin N(p^*) | V^n \in \Lambda) \leq 0$.
But trivially $\liminf_{n\to\infty} \Pr(P \notin N(p^*) | V^n \in \Lambda) \geq 0$, so that $\lim_{n\to\infty} \Pr(P \notin N(p^*) | V^n \in \Lambda) = 0$. ∎

Theorem 5.1 states that when $\nu(p)$ and p^* are unique (i.e. A1∩A2 holds), the posterior distribution when given the information that the empirical distribution lies in Λ degenerates, as the sample size grows, to a distribution concentrated at p^*. Moreover this p^* being the closest point to Λ in Ω depends only on the geometry of the sets Λ and Ω and is otherwise independent of the prior $\mathcal{P}$. An interpretation is that a decision maker who initially believes Ω to contain the possible distributions of the superpopulation should, in light of strong empirical evidence that the distribution is in Λ, adopt the view that P equals p^*.

Of course, the probability of finding the empirical distribution in Λ (when Λ and Ω do not meet) goes to zero as the sample size diverges to infinity. However for sufficiently large finite values of the sample or population size n, P will, with arbitrarily high probability, be near p^*. Hence the theorem entails an observable prediction. If one were to generate a series of large populations as samples from superpopulations with distribution P chosen in accordance with $\mathcal{P}$ over Ω, then for a population whose distribution lay in Λ, it is nearly certain to have been drawn from a superpopulation with distribution near p^*.

Using the uniqueness results of section 4, we derive from Theorem 5.1 the following result, which expresses explicitly sufficient conditions under which $\mathcal{P}^n \stackrel{d}{\Longrightarrow} p^*$.

Corollary 5.2 *Let the prior $\mathcal{P}$ be an absolutely continuous probability measure with support $\Omega \subset S$ where Ω is a closed convex set. Let $\Lambda \subset S$ be a closed convex body with $\Lambda \cap \Omega = \emptyset$. If either Ω or Λ is strictly convex or a singleton, or if $S \subset R^2$, then*

$$\mathcal{P}^n \stackrel{d}{\Longrightarrow} p^* \text{ as } n \to \infty$$

in the sense that $\mathcal{P}^n(N(p^*)) \to 1$ as $n \to \infty$ for all open neighborhoods $N(p^*)$ of p^*. ∎

From the above results, in the presence of convexity or uniqueness conditions, the uncertainty about the true distribution P as expressed by $\mathcal{P}$ unambiguously results in certainty that P is p^* in the face of incomplete but conflicting information on a large sample of observations. This convergence to a known distribution in the presence of only partial sample information and a general nonparametric prior is quite striking.

We turn now to consider the behavior of the empirical or population distribution V^n as n increases. We begin by demonstrating in Lemma 5.3 below that V^n converges in conditional probability to $v(p)$, when given that $P = p$.

Lemma 5.3 *Let $\Lambda, \Omega \subset S$ with Ω closed and Λ a closed body. Assume that for all $p \in \Omega$, $v(p) = \arg\min_{v \in \Lambda} I(v,p)$ is unique. Then*

$$\Pr(V^n \notin N(\nu(p)) | V^n \in \Lambda, P = p) \to 0$$

uniformly over $p \in \Omega$ as $n \to \infty$, where $N(\nu(p))$ is an open neighborhood of $\nu(p)$.

Proof: Let $\epsilon > 0$. If $\tilde{N}(\nu(p)) = \Lambda - N(\nu(p)) = \emptyset$, then

$$\Pr(V^n \notin N(\nu(p)) | V^n \in \Lambda, P = p) = 0 < \epsilon$$

for all n.

Suppose then $\tilde{N}(\nu(p)) \neq \emptyset$. Let $\delta = I(\tilde{N}(\nu(p))) - I(\Lambda, p) > 0$ since $\nu(p)$ is unique. By Lemma 3.3, we can choose N_1, so that for all $n > N_1$ and all $p \in \Omega$,

$$\frac{\log \Pr(V^n \in \tilde{N}(\nu(p)) | P = p)}{n} < -I(\tilde{N}(\nu(p)), p) + \delta/3$$

and choose N_2 so that for all $n > N_2$ and all $p \in \Omega$,

$$\frac{\log \Pr(V^n \in \Lambda) | P = p)}{n} > -I(\Lambda, p) - \delta/3.$$

Now choose N_3 so that

$$\log \epsilon^{1/n} > -\delta/3$$

for all $n > N_3$.

Let $N = \max(N_1, N_2, N_3)$. Then for all $n > N$ and all $p \in \Omega$,

$$\frac{\log \Pr(V^n \in \tilde{N}(\nu(p)) | P = p)}{n} - \frac{\log \Pr(V^n \in \Lambda | P = p)}{n} <$$
$$-I(\tilde{N}(\nu(p)), p) + I(\Lambda, p) + (2/3)\delta = -\delta/3 < \log \epsilon^{1/n}.$$

Hence $\Pr(V^n \notin N(\nu(p)) | V^n \in \Lambda, P = p) = \Pr(V^n \in \tilde{N}(\nu(p)) | P = p) / \Pr(V^n \in \Lambda | P = p) < \epsilon$ for all $n > N$ and all $p \in \Omega$. ∎

Theorem 5.4 (Convergence of the Empirical Distribution) *Let P, Ω, and Λ be as in Theorem 5.1. Consider the conditional sample distribution $\mathcal{V}^n(B) = \Pr(V^n \in B | V^n \in \Lambda)$ for all Borel sets $B \subset S$. Then*

$$\mathcal{V}^n \stackrel{d}{\Longrightarrow} \nu^* \text{ as } n \to \infty$$

in the sense that $\mathcal{V}^n(N(\nu^*)) \longrightarrow 1$ as $n \to \infty$ for all neighborhoods $N(\nu^*)$ of ν^*.

Proof: Let $\gamma > 0$. Let $\epsilon > 0$ be small enough so that $\Lambda_\epsilon(\nu^*) \subset N(\nu^*)$. Choose $\delta > 0$ so that $\nu(p) \in \Lambda_{\epsilon/2}(\nu^*)$ for all $p \in \Omega_\delta(p^*)$, which is possible by Lemma 4.6.
Now choose N_1 so that

$$\Pr(V^n \notin \Lambda_{\epsilon/2}(\nu(p)) | V^n \in \Lambda, P = p) < \gamma/2$$

for all $n > N_1$ for all $p \in \Omega_\delta(p^*)$ which is possible by Lemma 5.3. Then

$$\Pr(V^n \notin \Lambda_\epsilon(\nu^*) | V^n \in \Lambda, P = p) < \gamma/2$$

for all $n > N_1$, and $p \in \Omega_\delta(p^*)$.
Now

$$\Pr(V^n \notin N(\nu^*) | V^n \in \Lambda) \leq \Pr(V^n \notin \Lambda_\epsilon(\nu^*) | V^n \in \Lambda)$$

$$= \int_{\Omega_\delta(p^*)} \Pr(V^n \notin \Lambda_\epsilon(\nu^*) | V^n \in \Lambda, P = p) d\mathcal{P}(p | V^n \in \Lambda)$$
$$+ \int_{\Omega - \Omega_\delta(p^*)} \Pr(V^n \notin \Lambda_\epsilon(\nu^*) | V^n \in \Lambda, P = p) d\mathcal{P}(p | V^n \in \Lambda). \quad (11)$$

The second term on the right hand side of (10) is bounded from above by

$$\int_{\Omega - \Omega_\delta(p^*)} d\mathcal{P}(p | V^n \in \Lambda) = \Pr(P \notin \Omega_\delta(p^*) | V^n \in \Lambda) \leq \gamma/2$$

for $n > N_2$ for some N_2 by Theorem 5.1. The first term on the RHS of (10) is bounded from above by

$$\int_{\Omega_\delta(p^*)} (\gamma/2) d\mathcal{P}(p | V^n \in \Lambda) = (\gamma/2) \Pr(P \in \Omega_\delta(p^*) | V^n \in \Lambda) \leq \gamma/2$$

for all $n > N_1$. Let $N = \max\{N_1, N_2\}$. Then $\Pr(V^n \notin N(\nu^*) | V^n \in \Lambda) < \gamma$ for $n > N$. Hence $\Pr(V^n \notin N(\nu^*) | V^n \in \Lambda) \longrightarrow 0$ as $n \to \infty$. ∎

When the uniqueness condition A1∩A2 holds, Theorem 5.4 states the following. Suppose we generate a large population by drawing a random sample from a superpopulation whose distribution P has been determined in accordance with a prior distribution P over Ω. If the distribution of that population exhibits characteristics that depart from those characteristic of the superpopulation (i.e. $V^n \in \Lambda$), then that population's distribution is almost certain to be nearly ν^*. There is a similar classical interpretation of this result for the case when $P = p^*$ is known. This special case extends the result of Vasicek [1980] and is related to that of Van Campenhout and Cover [1981].

Once again from Lemma 5.3, we obtain the following corollary.

Corollary 5.5 *Let P, Ω, and Λ be as in Corollary 5.2. Then*

$$\mathcal{V}^n \stackrel{d}{\Longrightarrow} \nu^* \text{ as } n \to \infty$$

in the sense that $\mathcal{V}^n(N(\nu^*)) \longrightarrow 1$ as $n \to \infty$ for all open neighborhoods $N(\nu^*)$ of ν^*. ∎

6 Discussion and Conclusion

If we adopt the viewpoint expressed in the introduction that P is the distribution for the superpopulation and V^n is the distribution of a large population of n members drawn from this superpopulation, then the following observation flows from Theorem 5.1 and 5.4. When given the partial information that the population exhibits aggregate properties not shared by the superpopulation (i.e., that its distribution lies in Λ), then it is nearly certain that its distribution is close to ν^*. Moreover, the unknown distribution P, known to lie in Ω, is nearly certain to be near p^*. It follows that in a problem of inference under conflicting and incomplete information, it is essential to identify which information relates to the population as opposed to the superpopulation. The reason is that because of the asymmetry of $I(q,p)$, ν^* and p^* are not invariant when Λ and Ω are interchanged. For example, if we let Λ be the set of distributions with mean at most u and Ω be the singleton p^*, then ν^* is given by that member of the discrete exponential family containing p^* that has mean u (see Sampson and Smith [1985]). In particular, if p^* is a binomial distribution, then ν^* is that binomial distribution with mean u.

On the other hand, if we assign Ω to be the set of distributions with mean at least u and Λ to a small neighborhood around the point ν^*, then p^* is no longer a member of the exponential families. In fact, p^* is given by

$$p_i^* = \frac{\nu_i^*}{\lambda(i-u)+1} \text{ for } i = 0,1,\ldots,m$$

where λ is the unique root of certain mth degree polynomial equation [see Appendix II].

In Sampson and Smith [1982, 1985], the information provided by "expert judgement" that the mean is at most u is ascribed to Λ so as to obtain the exponential family estimate ν^* for the "true" distribution. Ω is set to the singleton $\{p^*\}$ where p^* is interpreted as the "decision maker's prior". From the results of this paper, this estimate is justifiable only under the interpretation that the decision maker's prior was in fact used to generate a population with mean at most u. Hence within this interpretation, the informal use of the Bayesian term "prior" by Sampson and Smith is rigorously justifiable. They also show that $\nu_0^* = u + o(u)$ independently of p^*, the so-called rare event approximation. Although ν^* in that context merely represented the closest point in Λ in relative entropy to p^*, Theorem 5.4 now provides a rigorously substantiated and empirically observable claim.

The duality expressed by ν^* and p^*, the former a classical and the latter a Bayesian viewpoint, is rich in paradoxes. For example, given two pieces of incomplete but conflicting information, which should play the role of Λ and which of Ω? At least within the model of this paper, the answer seems to pivot on a) which information is primary (Ω) and which secondary (Λ) and b) for which population is the estimate desired, the "population" or

"superpopulation"? The estimate ν^* is appropriate for the former while p^* is appropriate for the latter.

References

1. Bahadur, R.R., **Some Limit Theorems in Statistics**, SIAM, Philadelphia, 1971.

2. Brockett, P.L., A. Charnes, and W.W. Cooper, "MDI Estimation Via Unconstrained Convex Programming," bf Communications in Statistics, B-9, 223-234, 1980.

3. Van Campenhout, J. and T. Cover, "Maximum Entropy and Conditional Probability," **IEEE Trans. on Information Theory**, IT-27, 483-489, 1981.

4. Cozzolino, J.M. and M.J. Zahner, "The Maximum-Entropy Distribution of the Future Market Price of a Stock," **Operations Research** 21, 1200-1211, 1973.

5. Hausdorff, F. **Set Theory**, Chelsea, N.Y., 1957.

6. Jaynes, E.T., "Information Theory and Statistical Mechanics," **Physical Review** 106, 620-630, 1956.

7. Johnson, R.W., "Axiomatic Characterization of the Directed Divergences and Their Linear Combinations," **IEEE Transactions on Information Theory**, IT-17, 641-650, 1979.

8. Rockafellar, R.T., **Convex Analysis**, Princeton, NJ, 1970.

9. Sampson, A.R. and R.L. Smith, "Assessing Risks Through the Determination of Rare Event Probabilities," **Operations Research** 30, 839-866, 1982.

10. Sampson, A.R. and R.L. Smith, "An Information Theory Model For the Evaluation of Circumstantial Evidence," **IEEE Trans. Systems , Man, and Cybernetics** SMC-15, 9-16, 1985.

11. Sanov, I.N., "On the Probability of Large Deviations of Random Variables," **IMS and AMS Translations of Probability and Statistics**(from Mat. Sbornik 42, 11-44), 1961.

12. Shore, J.E. and R.W. Johnson, "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy,"**IEEE Transactions on Information Theory**, IT-26, 26-37, 1980.

13. Thomas, M.U., "A Generalized Maximum Entropy Principle," **Operations Research** 27, 1188-1195, 1979.

14. Vasicek, O.A. "A Conditional Law of Large Numbers," **Annals of Probability** 8. 142-147, 1980.

15. Wilson, A.G., **Entropy in Urban and Regional Modeling**, Pion Limited, London, 1970.

Appendix I

Properties of the Relative Entropy Functional:

Let p, q be probability vectors in

$$S = \{x \in R^n | \sum_{i=0}^{m} x_i = 1, x_i > 0, i = 0, 1, \ldots, m\}$$

The relative entropy functional $I : S \times S \longrightarrow R$ is defined by

$$I(v, p) = \sum_{i=0}^{m} v_i \log(v_i/p_i).$$

We also denote by R_+^{m+1} the positive orthant: $\{x \in R^{m+1} : x > 0\}$.

(a) $I(v, p)$ is (jointly) convex on $R_+^{m+1} \times R_+^{m+1}$.

(b) $I(v, \cdot)$ is strictly convex on R_+^{m+1} $(\forall v \in R_+^{m+1})$ and

$I(\cdot, p)$ is strictly convex on R_+^{m+1} for all $p \in R_+^{m+1}$.

(c) $I(v, p) \geq 0 (\forall p \in S, v \in S)$ with equality if and only if $p = v$.

(d) I is continuous on $R_+^{m+1} \times R_+^{m+1}$.

(e) Let Λ be a convex subset of R^{m+1}. Then the function $I(\Lambda, p) = \inf_{v \in \Lambda} I(v, p)$ is a convex function of p on R_+^{m+1}.

Proof:

(a) The two variables function

$$f(v_i, p_i) = v_i \log(v_i/p_i)$$

is convex since its Hessian

$$H = \begin{bmatrix} 1/v_i & -1/p_i \\ -1/p_i & v_i/p_i^2 \end{bmatrix}$$

is clearly positive semi-definite. Indeed, $\forall x = (x_1, x_2) \in R^2$, we have $x^T H x = (\frac{x_1}{\sqrt{v_i}} - \frac{x_2\sqrt{v_i}}{p_i})^2 \geq 0$. But $I(p, q) = \sum_{i=0}^{m} f(v_i, p_i)$ and hence is convex.

(b) $f(\cdot, p_i)$ is strictly convex on R_+ for $p_i > 0$ and $f(q_i, \cdot)$ is strictly convex on R_+ for $q_i > 0$ and hence (b) follows.

(c) Applying the gradient inequality to the strictly convex function $g(x) = I(x,p)$, we obtain

$$I(p,v) \geq I(p,p) + (v-p)^T \bigtriangledown I(p,p) \text{ with equality if and only if } p = v,$$

i.e.,

$$I(p,v) \geq 0 + \sum(v_i - p_i) \cdot 1 = 0 \text{ since } p, v \in S.$$

(d) This follows from the general fact that a convex function is continuous in the relative interior of its domain. Here ri (dom I) = $R_+^{m+1} \times R_+^{m+1}$.

(e) The pointwise infimum of a jointly convex function over a convex set is a convex function. (See e.g. [Rockafellar [1970, p. 38-39]). QED

Appendix II

Let $\Omega = \{p \in S | \sum_{i=0}^{m} ip_i = u\}$ and $\Lambda = \{v^*\}$. The closest point $p^* \in \Omega$ to v^* is the optimal solution of

$$\min_{p \in R_+^{m+1}} \{\sum_{i=0}^{m} v_i^* \log(v_i^*/p_i) = -\sum_{i=0}^{m} v_i^* \log p_i + \text{constant}\}$$

subject to

(i) $\sum_{i=0}^{m} ip_i = u$

(ii) $\sum_{i=0}^{m} p_i = 1.$

Let $\lambda \in R$ be the Lagrange multiplier of constraint (i) and $\mu \in R$ the Lagrange multiplier of constraint (ii). The optimality condition for p^* are then (i), (ii), and

(iii) $v_i^*/p_i^* - i\lambda - \mu = 0 \; i = 0, 1, \ldots, m.$

The solution of (i)-(iii) is

$$p_i^* = \frac{v_i^*}{\lambda(i-u)+1}$$

where λ is the unique solution of

(iv) $\sum_{i=0}^{m} \frac{v_i^*}{\lambda(i-u)+1} = 1.$

We note that (iv) is equivalent to the mth degree *polynomial* equation

$$\sum_{k=0}^{m} a_k \lambda^k = 0,$$

where the coefficient a_k is given by

$$a_k = \sum_{i_0} \sum_{i_1} \cdots \sum_{i_k} (i_j - u)(\sum_{j=0}^{k} v_{i_j}^*)$$

where $i_j \in \{0, 1, \ldots, m\}, i_j \neq i_{j+1}, j = 0, 1, \ldots, k-1.$

As a specific example let $m = 2$, and $u = 1$. Then

$$\lambda = \frac{v_2^* - v_0^*}{1 - v_1^*}, \; p_0^* = p_2^* = \frac{1 - v_1^*}{2}, \; p_1^* = v_1^*.$$