A Weak Law of Large Numbers for Rare Events

by

Donald E. Brown
Robert L. Smith

Technical Report 86-4
February 1986

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

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

A Weak Law of Large Numbers for Rare Events

Donald E. Brown
Robert L. Smith

Technical Report 86-4
February 1986

Abstract

We show that the empirical distribution associated with a discrete probability distribution p, when constrained to lie within a convex information set Λ, will as the number of trials increases become arbitrarily close with arbitrarily high probability to the distribution that minimizes the relative entropy between p and Λ.

A Weak Law of Large Numbers for Rare Events

Statistical inference is traditionally concerned with using data and probability models to derive conclusions about a practical problem with inherent variability. Classical statistical inference uses only the data and the sampling function to derive conclusions about the sampled population. However, a large class of problems involves inference with data in the form of deterministic constraints on the underlying probability model. Typically the constraints do not uniquely determine the unknown distribution.

A common approach to problems of this kind has been to use an information theoretic procedure known as relative entropy minimization. A special case of the relative entropy minimization procedure, known as entropy maximization, has been used in a wide variety of applications; examples include reliability (Tribus [1969]), urban modeling (Wilson [1970]), stock market pricing (Lozzolino and Zahner [1973]), oil spill damage assessment (Thomas [1979]), and statistical mechanics (Jaynes [1956]). Applications of the more general relative entropy minimization approach can be found in the areas of statistics (Kullback [1959]), statistical mechanics (Hobson [1971]), legal inference (Sampson and Smith [1984]), and risk assessment (Sampson and Smith [1982]). We provide in this paper a relative frequency interpretation of the relative entropy minimization procedure that lends empirical justification to the approach.

1. Relative Entropy Minimization as an Inference Procedure

The maximum entropy principle proposed by Jaynes [1956] was intended to be employed as a user invariant method of assigning probabilities based on testable information. Information in this context consisted of inequality constraints on the unknown distribution.

Relative entropy minimization is a more general inference procedure which admits an initial or prior distribution, in addition to the constraint information.

More formally, let $p = (p_0, p_1, \ldots, p_m) > 0$ be the prior probability distribution expressed as a probability mass function and let Λ, the information constraint, be a closed convex subset of the simplex S of all non-degenerate m+1 dimensional discrete distributions where $S = \{q | \sum_{i=0}^{m} q_i = 1, q_1 > 0 \text{ for } i = 0, 1, 2, \ldots, m\}$. Then the principle of minimum relative entropy prescribes choosing that q^* which minimizes the relative entropy subject to satisfying the constraint Λ, that is

$$q^* = \underset{q \varepsilon \Lambda}{\operatorname{argmin}} I(q, p)$$

where $I(q, p) = \sum_{i=0}^{m} q_i \ln (q_i/p_i)$. We are reduced to entropy maximization when $p_i = \frac{1}{m+1}$ for all i is the uniform distribution. $I(q, p)$ is also known as the cross-entropy (Shore and Johnson [1980]) or the Kullback-Leibler information discrimination (Kullback and Leibler [1951]).

Justifications for using relative entropy minimization in an inference procedure have relied on axioms of information (Hobson [1969], Hobson and Cheung [1973], and Johnson [1979]), or axioms of inference (Shore and Johnson [1980]).

2. A Correspondence Property for Relative Entropy Minimization

One of the principal justifications for the use of entropy maximization as an inference procedure was provided by Jaynes [1968]. He demonstrated a correspondence between the maximum entropy distribution and the most likely outcome in repeated trials of a random experiment.

In particular, let $a_0, a_1, \ldots, a_m$ be the possible outcomes of an experiment where each outcome is equally likely to occur. Suppose now we repeat the experiment n times and observe the number of times N_i that outcome i occurred for $i = 0, 1, \ldots, m$. Then Jaynes effectively showed that

$$\ln P(N_0 = n_0, N_1 = n_1, \ldots, N_m = n_m) = K_n \left(\sum_{i=0}^{m} \left(\frac{n_i}{n} \ln \frac{n_i}{n}\right) + C_m + \varepsilon_n \right)$$

where $K_n < 0$ and $C_m > 0$ are constants depending only on n and m respectively and $\varepsilon_n \to 0$ as $n \to \infty$ for fixed m.

From this result, Jaynes concluded the correspondence property that the probability distribution which maximizes the entropy is identical to the frequency distribution which can be realized in the greatest number of ways. We will extend and strengthen this correspondence property to the general minimum relative entropy procedure.

Suppose now that $a_0, a_1, \ldots, a_m$ are the possible outcomes within an experiment for which outcome i occurs with probability p_i for $i = 0, 1, \ldots, m$. Let $V_i^n = N_i/n$ be the relative frequency of occurrence of outcome i in n repeated independent trials of the experiment. We refer to $V^n = (V_0^n, V_1^n, \ldots, V_m^n)$ as the empirical distribution based on n trials.

The first lemma is due to Sanov [1961].

Lemma 1: For any $\nu \in S$, $P(V^n = \nu) = e^{-n(I(\nu,p) + O(\ln n/n))}$ for all n with $n\nu$ integer where $O(\ln n/n)$ depends only on m and is independent of ν and p.

Proof: By Stirling's approximation (Knuth [1976], p. 111) $n! = \sqrt{2\pi n}\,(n/e)^n\,(1 + O(1/n))$, so that $\ln n! = n(\ln n - 1) + O(\ln n)$. Suppose that $n\nu_i$ is integer for all $i = 0, 1, \ldots, m$. Then $P(V^n = \nu) = (n!/(n\nu_0!\, n\nu_1! \ldots n\nu_m!))\, p_0^{n\nu_0}\, p_1^{n\nu_1} \ldots p_m^{n\nu_m}$, and therefore $\ln P(V^n = \nu) = \ln n! - \sum_{i=0}^{m} \ln (n\nu_i!) + \sum_{i=0}^{m} n\nu_i \ln p_i = n(\ln n - 1) + O(\ln n)$

$$- \sum_{i=0}^{m} n\nu_i (\ln n\nu_i - 1) + \sum_{i=0}^{m} O(\ln (n\nu_i)) + \sum_{i=0}^{m} n\nu_i \ln p_i =$$

$$- \sum_{i=0}^{m} n\nu_i \ln \nu_i + \sum_{i=0}^{m} n\nu_i \ln p_i + O(\ln n)$$ where $O(\ln n)$ depends only on m since $n\nu_i \leq n$ for all i. Hence $\ln P(V^n = \nu)$

$$= - n \sum_{i=0}^{m} \nu_i \ln \nu_i/p_i + O(\ln n).$$

We have then $P(V^n = \nu) = e^{-n(\sum_{i=0}^{m} \nu_i \ln \nu_i/p_i + O(\ln n/n))}$. ■

Lemma 2: Let $\Lambda_\varepsilon = S_\varepsilon(\nu^*) \cap \Lambda$ be an ε-neighborhood around $\nu^* = \underset{\nu \in \Lambda}{\operatorname{argmin}}\, I(\nu, p)$ where $p \notin \Lambda$ and $S_\varepsilon(\nu^*) = \{\nu \mid |\nu_i - \nu_i^*| < \varepsilon \text{ for } i = 0, 1, 2, \ldots, m\}$. Then for all $\varepsilon > 0$, $\lim_{n\to\infty} P(V^n \in \Lambda_\varepsilon | V^n \in \Lambda) = 1$.

Proof: Let $\bar{\Lambda}_\varepsilon = \Lambda - \Lambda_\varepsilon$. Then $P(V^n \in \bar{\Lambda}_\varepsilon)$

$$= \sum_{\substack{\nu\in\bar{\Lambda}_\varepsilon \\ \text{with } n\nu \text{ integer}}} P(V^n = \nu) = \sum_{\substack{\nu\in\bar{\Lambda}_\varepsilon \\ \text{with } n\nu \text{ integer}}} e^{-n(I(\nu,\, p) + O(\ln n/n))}$$

$$\leq N(\bar{\Lambda}_\varepsilon)\, e^{-n(I(\nu_\varepsilon,\, p) + O(\ln n/n))} \text{ where}$$

$I(\nu_\varepsilon, p) = \min_{\nu\in\bar{\Lambda}_\varepsilon} I(\nu, p)$ and $N(T)$ is the number of points $\nu \in T \subseteq S$ with $n\nu$ integer. Now $N(\bar{\Lambda}_\varepsilon) \leq N(S) \leq \binom{n+m}{m} = O(e^{m \ln n})$ for fixed m where $S = \{\nu \mid \sum_{i=0}^{m} \nu_i = 1,\ \nu_i > 0 \text{ for } i = 0, 1, 2, \ldots, m\}$. Hence $P(V^n \in \bar{\Lambda}_\varepsilon) \leq e^{-n(I(\nu_\varepsilon,\, p) + O(\ln n/n))}$. On the other hand, by the continuity of $I(\nu, p)$ in ν over S, we can choose $\varepsilon > \varepsilon' > 0$ small enough so that $I(\nu, p) < I(\nu_\varepsilon, p)$ for all ν in the closure of $\Lambda_{\varepsilon'}$. Let ν_ε^* be a point in the closure of $\Lambda_{\varepsilon'}$ such that $I(\nu_\varepsilon^*, p) = \max I(\nu, p)$ over all ν in the closure of $\Lambda_{\varepsilon'}$. Clearly, for all $n \geq N_\varepsilon$ for some N_ε, there is a $\nu \in \Lambda_{\varepsilon'}$ with $n\nu$ integer. Let ν_ε^n be any such point. Then for all $n \geq N_\varepsilon$, $P(V^n \in \Lambda) \geq P(V^n = \nu_\varepsilon^n) = e^{-n(I(\nu_\varepsilon^n,\, p) + O(\ln n/n))} \geq e^{-n(I(\nu_\varepsilon^*,\, p) + O(\ln n/n))}$. Hence for all $n \geq N_\varepsilon$, we have

$$P(V^n \in \bar{\Lambda}_\varepsilon \mid V^n \in \Lambda) = P(V^n \in \bar{\Lambda}_\varepsilon)/P(V^n \in \Lambda)$$

$$\leq e^{-n(I(\nu_\varepsilon,\, p) - I(\nu_\varepsilon^*,\, p) + O(\ln n/n))} \to 0$$

as $n \to \infty$ since $I(\nu_\varepsilon, p) > I(\nu_\varepsilon^*, p)$. ■

From Lemma 2, we may easily infer the following theorem.

Theorem (Weak Law): Let Λ be a closed convex subset of $S = \{q \mid \sum_{i=0}^{m} q_i = 1,\ q_i > 0$ for $i = 0, 1, 2, \ldots, m\}$ and $p \notin \Lambda$ be a point in S. Let V_i^n be the relative frequency of occurrence of outcome i in n independent trials of an experiment that results in outcome i with probability p_i for $i = 0, 1, 2, \ldots, m$.

Set $\nu^* = \operatorname{argmin}_{\nu \in \Lambda} I(\nu, p)$. Then for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|V_i^n - \nu_i^*| \geq \varepsilon \mid V^n \in \Lambda) = 0$$

Proof: From Lemma 2, we have $\lim_{n\to\infty} P(V^n \notin \Lambda_\varepsilon \mid V^n \in \Lambda) = 0$ and hence our result. ■

Proof: Let $\bar{\Lambda}_\varepsilon = \Lambda - \Lambda_\varepsilon$. Then $P(V^n \in \bar{\Lambda}_\varepsilon)$

$$= \sum_{\substack{\nu \in \bar{\Lambda}_\varepsilon \\ \text{with } n\nu \text{ integer}}} P(V^n = \nu) = \sum_{\substack{\nu \in \bar{\Lambda}_\varepsilon \\ \text{with } n\nu \text{ integer}}} e^{-n(I(\nu, p) + O(\ln n/n))}$$

$$\leq N(\bar{\Lambda}_\varepsilon)\, e^{-n(I(\nu_\varepsilon, p) + O(\ln n/n))} \text{ where}$$

$I(\nu_\varepsilon, p) = \min_{\nu \in \bar{\Lambda}_\varepsilon} I(\nu, p)$ and $N(T)$ is the number of points $\nu \in T \subseteq S$ with $n\nu$ integer. Now $N(\bar{\Lambda}_\varepsilon) \leq N(S) \leq \binom{n+m}{m} = O(e^{m \ln n})$ for fixed m where $S = \{\nu | \sum_{i=0}^{m} \nu_i = 1, \nu_i > 0 \text{ for } i = 0, 1, 2, \ldots, m\}$. Hence $P(V^n \in \bar{\Lambda}_\varepsilon) \leq e^{-n(I(\nu_\varepsilon, p) + O(\ln n/n))}$. On the other hand, by the continuity of $I(\nu, p)$ in ν over S, we can choose $\varepsilon > \varepsilon' > 0$ small enough so that $I(\nu, p) < I(\nu_\varepsilon, p)$ for all ν in the closure of $\Lambda_{\varepsilon'}$. Let ν_ε^* be a point in the closure of $\Lambda_{\varepsilon'}$ such that $I(\nu_\varepsilon^*, p) = \max I(\nu, p)$ over all ν in the closure of $\Lambda_{\varepsilon'}$. Clearly, for all $n \geq N_\varepsilon$ for some N_ε, there is a $\nu \in \Lambda_{\varepsilon'}$ with $n\nu$ integer. Let ν_ε^n be any such point. Then for all $n \geq N_\varepsilon$, $P(V^n \in \Lambda) \geq P(V^n = \nu_\varepsilon^n) = e^{n(I(\nu_\varepsilon^n, p) + O(\ln n/n))} \leq e^{-n(I(\nu_\varepsilon^*, p) + O(\ln n/n))}$. Hence for all $n \geq N_\varepsilon$, we have

$$P(V^n \in \bar{\Lambda}_\varepsilon | V^n \in \Lambda) = P(V^n \in \bar{\Lambda}_\varepsilon)/P(V^n \in \Lambda)$$
$$\leq e^{-n(I(\nu_\varepsilon, p) - I(\nu_\varepsilon^*, p) + O(\ln n/n))} \to 0$$

as $n \to \infty$ since $I(\nu_\varepsilon, p) > I(\nu_\varepsilon^*, p)$. ■

From Lemma 2, we may easily infer the following theorem.

Theorem (Weak Law): Let Λ be a closed convex subset of $S = \{q | \sum_{i=0}^{m} q_i = 1, q_i > 0$ for $i = 0, 1, 2, \ldots, m\}$ and $p \notin \Lambda$ be a point in S. Let V_i^n be the relative frequency of occurrence of outcome i in n independent trials of an experiment that results in outcome i with probability p_i for $i = 0, 1, 2, \ldots, m$.

Set $\nu^* = \operatorname{argmin}_{\nu \in \Lambda} I(\nu, p)$. Then for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|V_i^n - \nu_i^*| \geq \varepsilon \,|V^n \in \Lambda) = 0$$

Proof: From Lemma 2, we have $\lim_{n\to\infty} P(V^n \notin \Lambda_\varepsilon \,|V^n \in \Lambda) = 0$ and hence our result. ■

The general correspondence property suggested by the Theorem above may be summarized by the statement that the minimum relative entropy distribution is identical to the empirical distribution that would be observed after a large number of trials. In a related result, Van Campenhout and Cover [1981] demonstrated that the conditional probability distribution of a single trial when given a fixed sample mean converged to the minimum relative entropy distribution as the number of trials grew large. We have shown that the relative frequency distribution over all trials would also converge to this same minimum relative entropy distribution.

References

Campenhout, J. Van and T. Cover, "Maximum entropy and conditional probability", IEEE Transactions on Information Theory, IT-27, pp. 483-489, 1981.

Cozzolino, J. M. and M. J. Zahner, "The maximum-entropy distribution of the future market price of a stock", Operations Research, Vol. 21, pp. 1200-1211, 1973.

Hobson, A., "A new theorem of information theory", Journal of Statistical Physics, Vol. 1, pp. 383-391, 1969.

Hobson, A. and B. K. Cheung, "A comparison of the Shannon and Kullback information measures", Journal of Statistical Physics, Vol. 7, pp. 301-310, 1973.

Jaynes, E. T., "Information theory and statistical mechanics", Physical Review, Vol. 106, pp. 620-630, 1956.

Jaynes, E. T., "Prior probabilistics", IEEE Transactions on Systems Science and Cybernetics, SSC-4, pp. 227-241, 1968.

Johnson, R. W., "Axiomatic characterization of the directed divergences and their linear combinations", IEEE Transactions on Information Theory, IT-17, pp. 641-650, 1979.

Knuth, Donald E., The Art of Computer Programming, Vol. 1, Second edition, Addison-Wesley, Reading, Massachusetts, 1976.

Kullback, S., Information Theory and Statistics, John Wiley & Sons, New York, 1959.

Kullback, S. and R. A. Leibler, "On information and sufficiency", Annals Math. Statist., Vol. 22, pp. 79-86, 1951.

Sampson, Allan R. and Robert L. Smith, "An information theory model for the evaluation of circumstantial evidence", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 15, pp. 9-16, 1984.

Sampson, Allan R. and Robert L. Smith, "Assessing risks through the determination of rare event probabilities", Operations Research, Vol. 30, pp. 839-866, 1982.

Sanov, I. N., "On the probability of large deviations of random variables", *IMS and AMS Translations of Probability and Statistics*, (From Mat. Sbornik 42, pp. 11-44), 1961.

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, pp. 26-37, 1980.

Thomas, M. U., "A generalized maximum entropy principle", *Operations Research*, Vol. 27, pp. 1188-1195, 1979.

Tribus, M., *Rational Descriptions, Decisions and Designs*, Pergamon, New York, 1969.

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