Information Theory and Martingales

Andrew R. Barron
University of Illinois

Abstract

Fundamental limit properties hold for conditional entropy

$$\lim\downarrow H(X_0 \mid X_1,\ldots,X_n) = H(X_0 \mid X_1,X_2,\ldots),$$

mutual information

$$\lim\uparrow I(X_0; X_1,\ldots,X_n) = I(X_0; X_1,X_2,\ldots),$$

and informational divergence (relative entropy)

$$\lim\uparrow D(P_n \,||\, Q_n) = D(P \,||\, Q),$$

where P and Q are probability measures on a measurable space (Ω, F) and P_n and Q_n are the restrictions to sigma-fields F_n satisfying $F_n \uparrow F$. Early contributors to results of this type include Dobrushin and Pinsker (who used a characterization of information quantities as a supremum of informations for discrete random variables) and Perez, J. Hajek, and Moy (who used martingale convergence theorems). Alternative proofs have been developed in recent years by S. Kullback and his colleagues (using a chain rule for informational divergence) and by Barron (using the dominated convergence theorem). In this talk we present a simplified proof based on the chain rule,

$$D(P_n \,||\, Q_n) - D(P_m \,||\, Q_m) = \int \rho_n (\log \rho_n/\rho_m) dQ.$$

for $n > m$, where $\rho_n = dP_n/Q_n$. The convergence of ρ_n in $L_1(Q)$ and of $\log \rho_n$ in $L_1(P)$ (when $D(P \,||\, Q) < \infty$) are established as consequences of this chain rule, without invoking the martingale convergence theorem.

A new proof of the martingale convergence theorem for positive martingales is an unexpected byproduct of the information-theoretic analysis.

Information Theory and Martingales

Andrew R. Barron
University of Illinois

November 1990

Summary

Fundamental limit properties hold for conditional entropy

$$\lim\downarrow H(X_0 \mid X_1,\ldots,X_n) = H(X_0 \mid X_1,X_2,\ldots), \tag{1}$$

mutual information

$$\lim\uparrow I(X_0; X_1,\ldots,X_n) = I(X_0; X_1,X_2,\ldots), \tag{2}$$

and informational divergence (relative entropy)

$$\lim\uparrow D(P_n \,||\, Q_n) = D(P \,||\, Q), \tag{3}$$

where P and Q are probability measures on a measurable space (Ω, F) and P_n and Q_n are the restrictions to sigma-fields F_n satisfying $F_n \uparrow F$. We focus our attention on the conclusion (3), since the first two conclusions can be derived as consequences of it. Proofs of result of this type are in [1-3] (based on a representation of the informational divergence as a supremum of discrete divergences for finite partitions of the measurable space), in [4-6] (based on the martingale convergence theorems), and in [7,8] (based on the dominated convergence theorem). An alternative proof technique has recently been developed by Kullback et. al. [9] using a chain rule for informational divergence. In this talk we present a simplified proof of (3) based on the chain rule and discuss the implications for the convergence of densities and information densities. The information-theoretic techniques also provide a new proof of the martingale convergence theorem for positive martingales.

The informational divergence is defined by $D(P \,||\, Q) = E_P \log dP/dQ$ if $P << Q$ and $D(P \,||\, Q) = \infty$ otherwise. A basic inequality that we use states that if $P << Q$ then

$$E \mid \log dP/dQ \mid \leq D(P \,||\, Q) + (2D(P \,||\, Q))^{1/2}, \tag{4}$$

where the expectation is with respect to P, see [3],[10,p.339].

Conclusion (3) may be proved as follows. It is trivially true if D_n is an unbounded sequence, so suppose this sequence is bounded. Let $\rho_n = dP_n/dQ_n$. Taking the expected value (with respect to P_n) on both sides of the identity $\log \rho_n - \log \rho_m = \log \rho_n/\rho_m$ for $n > m$ yields the chain rule

$$D_n - D_m = \int \rho_n (\log \rho_n/\rho_m) dQ. \tag{5}$$

We have that D_n is increasing and hence convergent, so by the Cauchy sequence property, $D_n - D_m$ tends to zero as $n \to \infty$ and then $m \to \infty$. This yields the convergence to zero of the relative entropy on the right side of equation (5). Analogous to (4), we have the inequality

$$\int \rho_n \mid \log \rho_n/\rho_m \mid dQ \leq \int \rho_n (\log \rho_n/\rho_m) dQ + (2 \int \rho_n (\log \rho_n/\rho_m) dQ)^{1/2}. \quad (6)$$

Thus $\log \rho_n$ is a Cauchy sequence in $L_1(P)$ and hence convergent in $L_1(P)$, so it remains to identify the limit to be $\log dP/dQ$. We use the inequality

$$\int \mid \rho_n - \rho_m \mid dQ \leq (2 \int \rho_n (\log \rho_n/\rho_m) dQ)^{1/2} \quad (7)$$

to conclude that ρ_n is convergent in $L_1(Q)$ and we let ρ denote the limit. It follows that $\lim \int_A \rho_n dQ = \int_A \rho dQ$ for all measurable sets A. Now for each A in $\bigcup_n F_n$, we have that $\lim \int_A \rho_n dQ = P(A)$, so $P(A)$ and $\int_A \rho dQ$ agree on a generating collection of sets and hence are the same measures. Thus $P << Q$, $\rho = dP/dQ$, and $\lim \rho_n = \rho$ in probability with respect to P. It follows that the $L_1(P)$ limit of $\log \rho_n$ must equal $\log dP/dQ$. This completes the proof of conclusion (3).

Almost sure convergence of the densities ρ_n follows from the L_1 convergence by application of the maximal inequality, for $\varepsilon > 0$,

$$Q\{\sup_{n \geq m} \mid \rho_n - \rho_m \mid > \varepsilon\} \leq (1/\varepsilon) \int \mid \rho - \rho_m \mid dQ. \quad (8)$$

A new proof of martingale convergence properties is an unexpected byproduct of the information-theoretic analysis. Let Y_n be a positive martingale with respect to a probability measure Q, adapted to a sequence of sigma-fields $F_n \uparrow F$. Suppose Y_n is $L \log L$-bounded, that is, $\sup_n \int Y_n \log Y_n dQ < \infty$. Then $\rho_n = Y_n/c$ (where $c = EY_n$) defines a sequence of probability density functions for which the chain rule (5) is satisfied. It then follows from inequality (7) and the Cauchy sequence criterion that ρ_n and hence also Y_n are convergent in $L_1(Q)$ and we let Y denote the martingale limit. (Note that this proof works without presuming the existence of a probability measure P, the restriction of which gives rise to the measures $P_n(A) = \int_A \rho_n dQ$ for $A \in F_n$, but such a measure may be defined as $P(A) = \int_A Y dQ/c$.) Almost sure convergence of the martingale follows from L_1 convergence, by application of the maximal inequality.

Next we show that an information-theoretic convergence proof can also be given for positive martingales which are L_1-bounded but not necessarily $L \log L$-bounded. Let X_n be a positive martingale with respect to P, and let $c = E X_n$. Then $E \log (1+X_n)$ is a positive decreasing sequence, bounded by $\log (1+c)$, and hence it converges to a finite constant. Thus by the Cauchy

sequence property, as $n \to \infty$ and then $m \to \infty$,

$$E \log \frac{1+X_m}{1+X_n} \to 0. \tag{9}$$

For $n \geq m$, let $Q_{m,n}$ be the measure defined to have the following density with respect to P,

$$\frac{dQ_{m,n}}{dP} = \frac{1+X_n}{1+X_m}. \tag{10}$$

By the martingale property $E(1+X_n)/(1+X_m) = 1$ for $n \geq m$, so $Q_{m,n}$ is a probability measure. Since the density in (10) is strictly positive, it follows that $P << Q_{m,n}$ with density $dP/dQ_{m,n} = (1+X_m)/(1+X_n)$. Consequently,

$$D(P \,||\, Q_{m,n}) = E \log \frac{1+X_m}{1+X_n}, \tag{11}$$

which tends to zero as $n \to \infty$ and then $m \to \infty$ by (9). Applying inequality (4), it follows that $\log(1+X_n)$ is a Cauchy sequence in $L_1(P)$ and hence this sequence is convergent. By the continuity of the logarithm, it follows that X_n converges in probability to a random variable X. If the martingale sequence is uniformly integrable, this implies the L_1 convergence and hence the almost sure convergence in the same way as above. Even if it is not uniformly integrable (such that L_1 convergence is not possible), almost sure convergence of X_n follows since it is equivalent to the almost sure convergence for every $r > 0$ of the uniformly integrable martingales X_{τ_n}, where $\tau = \inf\{k : X_k > r\}$ and $\tau_n = \min\{n, \tau\}$.

We conclude by pointing out that the argument provides an information-theoretic proof of the convergence of the conditional densities $p(X_0 | X_1, X_2, \ldots, X_n)$ that arise in an examination of the conditional entropy sequence (1). These are the density sequences which arise most naturally in traditional information theory. One application of these results is to the sandwich proof of the Shannon-McMillan-Breiman theorem and its generalizations [11]. The proof of (1) gives a direct information-theoretic argument that the sandwich gap in [11] tends to zero. Thus the Shannon-McMillan-Breiman theorem can be proven using only the ergodic theorem and elementary information-theoretic considerations.

References

[1] Dobrushin, R. L. "General formulation of Shannon's main theorem in information theory," (Russian) *Usp. Mat. Nauk.*, **14**, p.3-104, 1959. (English Translation) *Transl. A.M.S., Ser. 2*, **33**, p.323-438, 1963.

[2] Dobrushin, R. L. "Passage to the limit under the information and entropy integrals," *Theory Probab. Appl.* (English Translation) **5**, p.25-32, 1960.

[3] Pinsker, M. S. *Information and Information Stability of Random Variables and Processes*, 1960. (English Translation) Holden-Day, San Francisco, 1964.

[4] Perez, A. "Notions generalisees d'incertitude d'entropy et d'information du point de vue de la theorie de martingales," *Trans. First Prague Conf. Inform. Theory, Statist. Decision Functions, Random Processes*, p.183-208, 1957.

[5] Hajek, J. "A property of J-divergence of marginal probability distributions," *Czechoslovak Math. J.* **8**, p.460-463, 1958.

[6] Moy, S. C. "Generalizations of Shannon-McMillan theorem," *Pacific J. Math.* **11** p.705-714, 1961.

[7] Barron, A. R. "The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.* **13**, p.1292-1303, 1985.

[8] Orey, S. "On the Shannon-Perez-Moy theorem," *Contemp. Math.*, **41**, p.319-327.

[9] Kullback, S., Keegel, J. C., and Kullback, J. H. *Topics in Statistical Information Theory*, Springer-Verlag, Berlin 1987.

[10] Barron, A. R. "Entropy and the central limit theorem" *Ann. Probab.*, **14**, p.336-342, 1986.

[11] Algoet, P. and Cover, T. "A sandwich proof of the Shannon-McMillan-Breiman theorem," *Ann. Probab.*, **16**, p.899-909, 1988.