Limits of Information, Markov Chains, and Projection

Andrew R. Barron
Yale University
Department of Statistics
P. O. Box 208290
New Haven, CT 06520
Andrew.Barron@yale.edu

For Presentation at ISIT-2000, Sorrento, Italy, June,26-30,2000.

September 1999

Abstract

The chain rule of information theory establishes monotonicity of information convergence and demonstrates that log densities form Cauchy sequences, convergent in L_1, proving information limits, Markov chain convergence to stationarity, and existence of information projections for convex sets of functions.

1 SUMMARY

Information Theory illuminates limit theorems of Probability. In this presentation we give a strengthened Markov Chain convergence theorem and show how it is a consequence of an information-theoretic analysis of sequences of information divergence. Similar analysis provides results for martingale convergence and for information projection.

Recall that for a pair of probability measures P and Q with densities p and q (with respect to a dominating measure) the information divergence is $D(P||Q) = E_P \log p(X)/q(X)$. In like manner we define $A(P||Q) = E_P|\log p(X)/q(X)|$, which is the expected absolute-value of the log-density ratio and $V(P,Q) = \int |p-q|$ which is the variation distance between P and Q. We use a Pinsker-type inequality (Barron 1985), $A \leq D + \sqrt{2D}$, deduced from $V \leq \sqrt{2}D$. Thus, for a sequence of probability measures, if the divergence D tends to zero, then so does A, that is, the log-densities converge in L_1.

In our examination of information limits, Markov chains, martingales, and existence of information projections we use one proof technique. The chain rule of information theory is used to deduce monotonicity of sequences of information divergence, to identify the difference sequence as a related information divergence, to deduce that the difference sequence tends to zero, and whence, using the Pinsker-type inequality, to deduce that certain log-densities form convergent Cauchy sequences in L_1, thereby ensuring the desired limit.

2 Markov Chains

Let $X_1, X_2, \ldots$ be a Markov chain with stationary transitions, let P_n denote the distribution after n steps, and suppose the chain has a unique stationary probability distribution P^*. Information theory in an analysis of the convergence of P_n to to P^* first arose in the work of Renyi (1960) and Kendall (1994). A key property is the decrease of the information divergence $D(P_n||P^*)$ as n increases (cf. Cover and Thomas 1990). Of particular note is the work of Fritz (1971) who used information inequalities to show (among other things) that, in a general state space setting, if the Markov chain is reversible, then finiteness of the sequence $D(P_n||P^*)$, implies total variation convergence of P_n to P^*. [Reversibility means that if the chain were started with the stationary probability distribution then the joint distribution of X_1, X_2 is the same as if the order were reversed.] Here I obtain the following strengthened Markov chain convergence result showing that the sequence $D(P_n||P^*)$ converges to zero.

Theorem: If $\{X_n\}$ is a reversible Markov chain with a unique stationary probability distribution P^*, then

$$\lim D(P_n||P^*) = 0$$

if and only if the sequence $D(P_n||P^*)$ is (eventually) finite.

Let $D_n = D(P_n||P^*)$. The heart of the proof is as follows, for $n > m$ the chain rule identifies the difference D_m - D_n as a divergence (between conditional distributions for X_m given X_n), which is nonnegative, establishing the monotonicity and convergence of the sequence $\{D_n\}$, so that by the Cauchy sequence property D_m - D_n tends to zero as n and m tend to infinity, and thus establishing, in light of the Pinsker-type inequality, that $A_{n,m} = E|\log p_m(X_m)/p^*(X_m) - \log p_n(X_n)/p^*(X_n)|$ tends to zero as n and m tends to infinity, so that $\log p_n(X_n)/p^*(X_n)$ is a Cauchy sequence in L_1 and hence convergent in L_1 (and in probability). By Fritz, for a reversible chain, we know that $p^*(X_n)/p_n(X_n)$ converges to one in probability. Thus $\log p_n(X_n)/p^*(X_n)$ which we have shown to be convergent in L_1 must have limit zero. Thus, $\lim D(P_n||P^*) = 0$.

3 Decreasing Information

Let P and Q be two probability measures on a measurable space and let $\mathcal{F}_n$ be a decreasing sequence of sigma-fields with limit (intersection) denoted $\mathcal{F}_\infty$. Let P_n and Q_n, respectively, denote the restrictions of P and Q to $\mathcal{F}_n$, and similarly let P_∞ and Q_∞ denote the restrictions to $\mathcal{F}_\infty$.

Theorem Limit of decreasing information: If the sequence $D(P_n||Q_n)$ is eventually finite, then

$$\lim D(P_n||Q_n) = D(P_\infty||Q_\infty).$$

The idea of the proof is the same. Let $D_n = D(P_n||Q_n)$. For simplicity, suppose $D(P||Q)$ is finite. Let ρ_n denote the probability density of P_n with respect to Q_n. For $n > m$ we have the chain rule identity $D_m - D_n = \int \rho_m \log \rho_m/\rho_n dQ$ which establishes the monotonicity, convergence, and, thereby, the Cauchy sequence properties of $\{D_n\}$, so that, via the Pinsker-type inequalities, both $\int |\rho_m - \rho_n| dQ$ and $A_{n,m} = E|\log \rho_m - \log \rho_m|$ tend to zero as n and m tends to infinity. That is, ρ_n is a Cauchy sequence in $L_1(Q)$ and $\log \rho_n$ is a Cauchy sequence in $L_1(P)$. Hence ρ_n is convergent in $L_1(Q)$ (we denote the limit as ρ_∞) and $\log \rho_n$ is convegent in $L_1(P)$ with limit $\log \rho_\infty$. Finally, considering sets A in $\mathcal{F}_\infty$, they are in $\mathcal{F}_n$ for all n, and hence for such sets we have $P(A) = \int_A \rho_n dQ$ for all n, and thus in view of the $L_1(Q)$ convergence $P(A) = \int_A \rho_\infty dQ$ for A in $\mathcal{F}_\infty$, that is, the limit ρ_∞ is identified as the density of the restrictions of P and Q to $\mathcal{F}_\infty$. Then from the $L_1(P)$ convergence of $\log \rho_n$ we deduce that $\lim D(P_n||Q_n) = D(P_\infty||Q_\infty)$ as claimed.

We remark that the Markov Chain convergence theorem is a consequence of this limit theorem for decreasing information. Indeed, let $\mathcal{F}_n$ be the σ-field generated by $X_n, X_{n+1}, \ldots$, let P be the distribution for the given random process $\{X_1, X_2, \ldots\}$, and let$Q$ be the corresponding distribution in which the process is stationary. These processes share the same conditional distributions for subsequent outcomes given X_n, so the divergence between the processes restricted to $\mathcal{F}_n$, addressed in the limit theorem for decreasing information, is the same as the divergence between the distributions of just X_n, addressed in the Markov Chain convergence theorem.

4 Increasing Information

One may also look at the case of an increasing sequence of sigma-fields $\mathcal{F}_n$ generating a limit sigma-field $\mathcal{F}$ on which we have probability measures P and Q. Again let P_n, Q_n denote the restrictions to $\mathcal{F}_n$. One may ask for the limit of $D(P_n||Q_n)$. Identification of the limit as $D(P||Q)$ is fundamental to information theory. This is a classical result by Kolmogorov and his colleagues, see Dobrushin (1959,1960) and Pinsker (1960) (who applies it to demonstrate the connection between information divergence and certain limits of discrete divergences and

to identify information rates for Gaussian processes) and various proofs are also in the work by Perez (1957), Moy (1961), and Barron (1985), who apply it to the identification of the conditional entropy limit in the asymptotic equipartition property. All of these appeal to martingale convergence theorems to obtain the conclusion.

Theorem Limit of increasing information:

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

.

Here we see that the conclusion follows more directly from the chain rule, without need to appeal to a martingale convergence theorem or domination inequalities. The proof is in essence the same as given above for the decreasing information case, with only slight modification. First note that by the chain rule D_n is increasing and never greater than $D(P||Q)$, so the conclusion is immediate if the sequence D_n is unbounded. If D_n is a bounded sequence, then since by monotonicity it is convergent, it is a Cauchy sequence so that $D_n - D_m$ tends to zero. The chain rule shows that this difference is equal to the divergence $\int \rho_n \log \rho_n/\rho_m dQ$ so by the Pinsker-type inequalities one obtains convergence of ρ_n in $L_1(Q)$ and $\log \rho_n$ in $L_1(P)$. Let ρ_∞ denote the $L_1(Q)$ limit of the ρ_n. Then for sets A in the union of the $\mathcal{F}_n$ one has $P(A) = \lim \int_A \rho_n dQ = \int_A \rho dQ$. Thus $P(A)$ and $\int_A \rho dQ$ agree on a generating collection of sets and hence are the same measures. Thus $P << Q$ with density ρ and it follows that the $L_1(P)$ limit of $\log \rho_n$ is $\log \rho$. Consequently, $\lim D(P_n||Q_n) = D(P||Q)$.

Martingale convergence is a consequence of the information-theoretic analysis. If Y_n is a nonnegative martingale with respect to a measure Q, then letting $\rho_n = Y_n/EY_n$ one obtains a probability density function for which if $\int \rho_n \log \rho_n dQ$ is a bounded sequence, one finds that ρ_n converges in $L_1(Q)$, leading to a measure P with density ρ_∞ with respect to Q and $\log \rho_n$ converges in $L_1(P)$. Truncation techniques permit analogous martingale converges results more generally for uniformly integrable martingales.

This proof of the increasing information limit, together with a demonstration of consequences for martingale convergence, was presented by the author at the 1990 Information Theory Symposium. The application of this technique to identify the information limit of Markov chains and the limit of decreasing information is new.

5 Information Projection

It may come as a surprise that questions of information projection are fundamentally of the same sort as discussed above.

Let C be a convex set of probability density functions $q(x)$ and let $p(X)$ be a given density, possibly outside of p. Loosely stated, the problem is one of minimizing either

$D(q||p)$ or $D(p||q)$ over choices of q in C. However, the issue arises that the class C might not admit a solution in C and one needs to understand the nature and existence of a limit q^* obtained by taking a sequence of q_n approaching the infimum of the divergence over choices of q in C. Let $D(C||p)$ and $D(p||C)$ denote the infimum of $D(q||p)$ and of $D(p||q)$, respectively, over choices of q in C.

The $D(C||p)$ case is thoroughly studied in Csiszar (1975,1984) and Topsoe (1979). Suppose that $D(C||p)$ is finite. A unique information projection Q^* with density q^* exists (though not necessarily in C) characterized by the property that for any sequence q_n in C with $D(q_n||p) \to D(C||p)$ we have $D(q_n||q^*) \to 0$ (thus q^* is also called the center of attraction of the convex set relative to p – Csiszar (1984) calls it a generalized information projection). It is also characterized by a Pythagorean-like inequality $D(q||p) \geq D(q||q^*)+D(C||p)$. The proof of existence of q^* is based on a decomposition of the increment in divergence $D(q_n||p)$ with increasing n together with a resulting demonstration of a Cauchy sequence property for the densities yielding the existence of the limit with the desired properties.

Distributions in C with divergence from p close to the infimum must be close to q^*. It is this property that leads in Csiszar (1984) to the conditional limit theorem, that is, the conclusion that the distribution Q_n of i.i.d. random variables, conditioned on the event that the empirical distribution is in the convex set C, must converge to Q^* in the sense that $D(Q_n||Q^*) \to 0$.

An interesting paradox of the theory is that though Q^* is a limit of members of the family Q_n in information divergence (and hence in total variation), one cannot overcome the difficulty of Q^* outside of C simply by replacing C by its closure $\bar{C}$. Indeed, as Csiszar (1984) shows the distances $D(\bar{C}||P)$ and $D(C||P)$ can be dramatically different.

Here we study information projection with the reversed order of arguments $D(p||C)$.

Let C be a convex set of nonnegative functions $q(x)$ of a random variable X (for some applications it is not relevant whether the functions q integrate to one). The problem is one of maximizing $E \log q(X)$ over choices of q in C (or, equivalently, minimizing $E \log p(X)/q(X)$). The problem might not admit a solution in C and one needs to understand the nature and existence of a limit q^* obtained by taking a sequence of q_n with $E \log q_n(X)$ approaching the supremum of $E \log q(X)$ over choices of q in C.

Examples to think of here are convex families of probability densities (e.g. all finite mixtures of Gaussian densities) for which minimization of $D(p||q)$ is of interest. For example, this minimization is the theoretical counterpart of maximum likelihood estimation and, as such, it arises in analysis of the M.L.E. In such cases the interest is to know what will be the limit of such estimates when the true density p is not in the convex family. This is the tactic taken in the thesis of Jonathan Li (1999) and the study of information projection given here is joint with him.

The work of Bell and Cover (1980,1988) shows that maximization of $E \log g(X)$ by a function g^* residing in C is equivalent to $Eg(X)/g^*(X) \leq 1$ for all g in C. They give

applications to growth rate optimal portfolio selection for convex combinations of a given set of stocks, where compactness of the simplex of portfolio vectors ensures that a maximizer exists. Later Algoet and Cover (1986) apply it to the asymptotic equipartition property and a counterpart A.E.P. for wealth sequences in portfolio selection). Kieffer (1999) takes the theory a step further. He presumes that the class of functions $\log g(X)$ is L_1 closed and uses a chain rule and Cauchy sequence argument to derive the existence of such a maximizer g^* in C building on the work of Bell and Cover, and he gives application to various extensions of ergodic theorems (including Kingman's subadditive ergodic theorem and asymptotic equipartition properties).

For general convex sets C of densities q we see that a information projection q^* exists (possibly outside of C) satisfying a Pythagorean-like identity as well as the Cover-Bell inequality. A function q^* is called the information projection if for every q_n with $D(p||q_n) \to D(p||C)$ we have $\log q_n \to \log q^*$ in $L_1(p)$.

Theorem: Let C be a convex set of probability densities and suppose $D(p||C)$ is finite. Then the information projection q^* exists, is unique, and satisfies the following properties. First, $D(p||q^*) = D(p||C)$; second, $c_q = Eq(X)/q^*(X) \leq 1$ for all $q \in C$; and third, for each q in C, defining a probability density $r = (pq/q^*)/c_q$, we have a Pythagorean-like inequality $D(p||q) \geq D(p||q^*) + D(p||r)$.

Here $D(p||r)$ is a disguised distance between q^* and q. As we will see it provides an upper bound for $E|\log q^*(X) - \log q(X)|$.

We will not give the whole proof here, focusing instead of the part that is in common with the information limit results presented above.

Consider a sequence q_n in C such that $D(p||q_n) \to D(p||C)$ and assume $D(p||q_n)$ is finite for each n. Take the first n of these $q_1, q_2, \ldots, q_n$ and let C_n be their convex hull, parameterized by the $n-$simplex, on which the divergence is a continuous function of the parameters of convex combination, so there is a minimizer $\tilde{q}_n$, which by Bell and Cover satisfies $Eq(X)/\tilde{q}_n(X) \leq 1$ for q in C_n.

Thus we have that $D_n = D(p||\tilde{q}_n)$ is a decreasing sequence converging to $D(p||C)$ and $c_{m,n} = E\tilde{q}_m(X)/\tilde{q}_n(X) \leq 1$ for all $n > m$. Now with $r_{m,n} = (p\tilde{q}_m/\tilde{q}_n)/c_{m,n}$, we find that the difference $D_m - D_n$ equals $D(p||r_{m,n}) + \log 1/c_{m,n}$ where both of these terms are non-negative. So by the Cauchy sequence property for D_n we deduce that both $D(p||r_{m,n})$ and $\log 1/c_{m,n}$ converge to zero as n and m tend to infinity. Consequently, $c_{m,n} \to 1$ and by the Pinsker-type inequality we have that $E|\log p(X)/r_{m,n}(X)|$ tends to zero, which means that $E|\log \tilde{q}_m(X)/\log \tilde{q}_n(X)|$ tends to zero. Thus $\log \tilde{q}_n(X)$ is a Cauchy sequence in $L_1(p)$, we denote its limit as $\log q^*(X)$, or, more to the point, $\log p(X)/\tilde{q}_n(X)$ is a Cauchy sequence with limit $\log p(X)/q^*(X)$ and $D(p||q^*) = \lim D(p||\tilde{q}_n) = D(p||C)$.

Now one may expand C to $C^* = \{\alpha q + (1-\alpha)q^* : q \in C, 0 \leq \alpha \leq 1\}$ which is a convex set in which one finds that $D(p||C^*) = D(p||C)$ and $c_q = E(q(X)/q^*(X) \leq 1$ for all q in C.

With $r = (pq/q^*)/c_q$ have the identity

$$D(p||q) - D(p||q^*) = D(p||r) + \log 1/c_q,$$

so with $c_q \leq 1$ we have the claimed Pythagorean-like inequality. Moreover, for any sequence of q's if $D(p||q)$ approaches $D(p||C)$, then both $D(p||r)$ and $\log 1/c_q$ approach zero. Consequently, c_q approaches 1 and, by the Pinsker-type inequality, $E|\log q(X) - \log q^*(X)|$ tends to zero. Thus q^* is the unique information projection.

Further properties of the information projection q^* are that if C is a family of probability densities then q^* integrates to not more than one and if also for some q_0 in C the family of functions q/q_0 is bounded, then q^* integrates to one.

In the case that C is a convex hull of a family Φ of densities satisfying the bounded ratio property, the information projection is shown in Li (1999) (c.f. Li and Barron (1999)) to be reached in the limit by an iterative sequence of optimizations of densities of the form $q_k = (1-\alpha)g_{k-1} + \alpha\phi$, with error $D(p||q_k) - D(p||C)$ approaching zero at rate $1/k$.

References

P. H. Algoet and T. M. Cover. A sandwich proof of the Shannon-McMillan-Breiman theorem. *Ann. Probab.* vol.16, pp.899-909, 1988.

A. R. Barron. The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.*, Vol.13, pp.1292-1303, 1985.

A. R. Barron. Entropy and the central limit theorem. *Ann. Probab.*, Vol.14, pp.336-342, 1986.

A. R. Barron. Information theory and martingales. Presented at the *1991 International Symposium on Information Theory.*

R. Bell and T. M. Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research* vol.5, pp.161-166, 1980.

R. Bell and T. M. Cover. Game-theoretic optimal portfolios. *Management Science* vol.34, pp.724-733, 1988.

T. M. Cover and J. A. Thomas. *Elements of Information Theory* Wiley, New York, 1991.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* vol.3, pp.146-158, 1975.

I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.* vol.12, pp.768-793, 1984.

J. Fritz. An information-theortetical proof of limit theorems for reversible Markov processes. *Trans. Sixth Prague Conf. on Inform. Theory, Statist. Decision Functions, Random Processes* Prague, Sept. 1971, Academia Publ., Czech. Acad. Science, 1973.

D. G. Kendall. *Information Theory and the limit theorem for Markov chains and processes with a countable infinity of states.* Ann. Inst. Stat. Math.} vol.15, pp.137-143, 1964.

J. Kieffer. An almost sure convergence theorem for sequences of random variables selected from log-convex sets. In *Almost everywhere convergence II*, p.151, Academic Press.

S. Kullback. A lower bound for discrimination in terms of variation, *IEEE Trans. Inform. Theory*, vol.13, pp.126-127, 1967.

S. Kullback, J. C. Keegel, and J. H. Kullback. *Topics in Statistical Information Theory.* Springer-Verlag, Berlin, 1980.

J. Q. Li. Estimation of Mixture Models. Ph.D Dissertation. Department of Statistics. Yale University. May 1999.

J. Q. Li and A. R. Barron. Mixture Density Estimation. To appear in *Proceedings Conference on Neural Information Processing Systems: Natural and Synthetic.* Denver, C0, December 1999.

S. C. Moy. Generalizations of Shannon-McMillan theorem. *Pacific J. Math.* vol.11, 705-714, 1961.

M. S. Pinsker. *Information and Information Stability of Random Variables*, Transl. by A. Feinstein. Holden-Day, San Francisco, 1964.

A. Rényi. On measures of entropy and information. In *Proc. Fourth Berkeley Symp. on Math. Statist. and Probab.*, vol.I, pp.547-561, 1960.

F. Topsoe. Information theoretical optimization techniques. *Kybernetika* vol.15, pp.8-27, 1979.