Limits of Information, Markov Chains, and Projection

Andrew R. Barron
Yale University, Statistics Dept.
Box 208290, New Haven, CT 06520
e-mail: Andrew.Barron@yale.edu

Abstract — **The chain rule of information shows that log densities form Cauchy sequences, convergent in L_1, proving information limits, Markov chain convergence, and existence of information projections.**

Let $D(P||Q) = E_P \log p(X)/q(X)$, $A = E|\log p(X)/q(X)|$, and $V = \int |p-q|$ be the information divergence, absolute information divergence, and total variation distance between probability measures P and Q with density functions p, q with respect to a dominating measure on a measurable space. The chain rule and the Pinsker-type inequality $A \leq D + \sqrt{2D}$, deduced from $V \leq \sqrt{2D}$ (which implies that if D tends to zero then so does V and A) allow one to deduce in various settings that log densities provide Cauchy sequences convergent in L_1, thereby establishing information limits including Markov chain convergence and information projections.

I. Markov chains

Let $\{X_n\}$ be Markov with stationary transition probability on a general state space and let P_n be the distribution of X_n.

Theorem 1. *Markov Chain Convergence.* If $\{X_n\}$ is a reversible Markov chain with a unique invariant probability distribution P^*, then $\lim D(P_n||P^*) = 0$ if and only if the sequence $D(P_n||P^*)$ is eventually finite.

Proof: Let $D_n = D(P_n||P^*)$. The chain rule gives $D_m - D_n$, for $n > m$, as a divergence (between conditional distributions for X_m given X_n), establishing monotonicity and convergence of D_n, so that $D_m - D_n \to 0$ as $n, m \to \infty$, and thus via the Pinsker-type inequality $E|\log p_m(X_m)/p^*(X_m) - \log p_n(X_n)/p^*(X_n)| \to 0$, so that $\log p_n(X_n)/p^*(X_n)$ is a Cauchy sequence, convergent in L_1. Fritz [4] used information inequalities for reversible chains to show the total variation convergence of P_n to P^*, so that $p^*(X_n)/p_n(X_n)$ converges to 1 in probability. Thus $\log p_n(X_n)/p^*(X_n)$, which we have shown to be convergent in L_1, must have L_1 limit equal to 0.

II. Information Limits

Let $\mathcal{F}_n$ be a monotone sequence of sigma-fields with limit $\mathcal{F}_\infty$. Let P_n and Q_n denote the restrictions of P and Q to $\mathcal{F}_n$, let ρ_n be the density of P_n with respect to Q_n, and let $D_n = D(P_n||Q_n)$ for $n = 1, 2, \ldots, \infty$.

Theorem 2. *Information Limit.* If $\mathcal{F}_n$ is decreasing or if $\mathcal{F}_n$ is increasing and $D(P_n||Q_n)$ is bounded, then $\log \rho_n \to \log \rho_\infty$ in $L_1(P)$ and $\lim_n D(P_n||Q_n) = D(P_\infty||Q_\infty)$.

Proof: In the case that $\mathcal{F}_n$ is decreasing, for $n > m$ we have $D_m - D_n = \int \rho_m \log \rho_m/\rho_n dQ$ establishing monotonicity, convergence, and, hence, the Cauchy sequence property, so that, via the Pinsker-type inequalities, both $\int |\rho_m - \rho_n| dQ$ and $E|\log \rho_m - \log \rho_n|$ tend to 0 as $n, m \to \infty$. Hence ρ_n is convergent in $L_1(Q)$ (denote the limit ρ_∞) and $\log \rho_n$ is convergent in $L_1(P)$ with limit $\log \rho_\infty$. Sets A in $\mathcal{F}_\infty$ are in $\mathcal{F}_n$ for all n with $P(A) = \int_A \rho_n dQ$, so by $L_1(Q)$ convergence, $P(A) = \int_A \rho_\infty dQ$, that is, the limit ρ_∞ is indeed the density between the restrictions of P and Q to $\mathcal{F}_\infty$. For the increasing case one proceeds in the same manner using the chain rule to extract Cauchy convergence of ρ_n in $L_1(Q)$ and $\log \rho_n$ in $L_1(P)$ and to identify the limit.

Theorem 2 implies Theorem 1 using the decreasing $\mathcal{F}_n$ generated by $\{X_n, X_{n+1}, \ldots\}$. The conclusion for the limit of increasing information is classical, see [1] and the references cited therein. Our analysis shows the convergence directly from the chain rule, without appeal to a martingale convergence theorem. The results for the limit of decreasing information and the information limit of Markov chains are new.

III. Information Projection

Demonstrating existence of information projections for convex sets of distributions uses similar techniques. 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 a convex set C. The set C might not admit a minimizer and one seeks a limit q^* obtained by sequences of q_n approaching the infimum. Topsoe [7], see also [3], resolves the $D(C||p)$ case. Here we state a result for the $D(p||C)$ case developed further in the Thesis of Li [6].

Theorem 3. *Information Projection.* Let C be convex and $D(p||C)$ finite. There exists a unique q^* (possibly outside of C) such that every sequence q_n with $D(p||q_n) \to D(p||C)$ has $\log q_n \to \log q^*$ in $L_1(p)$. Thus $D(p||q^*) = D(p||C)$. For all q in C, $c_q = E_p q(X)/q^*(X) \leq 1$ and, defining the density $r = (pq/q^*)/c_q$, we have the Pythagorean-like inequality $D(p||q) \geq D(p||q^*) + D(p||r)$, where via the Pinsker-type inequality $D(p||r)$ controls the $L_1(P)$ distance between $\log q$ and $\log q^*$. Furthermore, if $\int q = 1$ for all q in C, then $\int q^* \leq 1$.

Previously, Bell and Cover [2] show characterizing properties if q^* is in C. Kieffer [5] shows if $\{\log q : q \in C\}$ is closed in $L_1(P)$, then there exists q^* satisfying the key properties.

The proof identifies a sequence q_n in C such that $D(p||q_n) \downarrow D(p||C)$ and $c_{m,n} = E q_m(X)/q_n(X) \leq 1$ for all $n > m$. With $r_{m,n} = (pq_m/q_n)/c_{m,n}$, one finds $D_m - D_n$ equals $D(p||r_{m,n}) + \log 1/c_{m,n}$, so by the Cauchy sequence property, $\log 1/c_{m,n}$, $D(p||r_{m,n})$ and hence $E|\log q_m(X)/\log q_n(X)|$ converge to 0 as $n, m \to \infty$. Thus $\log q_n$ is a Cauchy sequence with limit denoted $\log q^*$ in $L_1(p)$. Further details are in [6].

References

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

[2] R. Bell and T. M. Cover. Competitive optimality of logarithmic investment. *Math. of Oper. Res.* vol. 5, pp. 161-166, 1980.

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

[4] J. Fritz. An information-theoretical proof of limit theorems for reversible Markov processes. *Trans. Sixth Prague Conf. on Inform. Theory, Stat. Dec. Func., Rand. Proc.* Czech. Acad. 1973.

[5] J. Kieffer. An almost sure convergence theorem for sequences of random variables selected from log-convex sets. In *Almost everywhere convergence II*, pp. 151-166, Academic Press, 1991.

[6] J. Q. Li. Estimation of Mixture Models. Yale Thesis, 1999.

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

0-7803-5857-0/00/$10.00 ©2000 IEEE.