Martingales and Information Divergence

Peter Harremoës
Mathematical Department, University of Copenhagen
Universitetsparken 5, DK-2100 Copenhagen O
Denmark

Abstract—A new maximal inequality for non-negative martingales is proved. It strengthens a well-known maximal inequality by Doob, and it is demonstrated that the stated inequality is tight. The inequality emphasizes the relation between martingales and information divergence. It implies pointwise convergence of $X \log X$ bounded martingales. A similar inequality holds for ergodic sequences. Relations to the Shannon-McMillan-Breiman theorem and Markov chains are mentioned.

Index Terms—Convergence, Individual Ergodic Theorem, information divergence, martingale, maximal inequality.

I. INTRODUCTION

The law of large numbers exists in different versions. The weak law of large numbers states that an empirical average is close to the theoretical mean with high probability for a large sample size. By use of large deviation bounds it is easy to prove the weak law of large numbers. The strong law of large numbers states that the empirical average converges to the theoretical mean with probability one. Large deviation bounds give an exponentially decreasing probability of a large deviation which implies the strong law of large numbers. These large deviation bounds are closely related to information theory. There are two important ways of generalizing the law of large numbers. One is in the direction of martingales. The other is in the direction of ergodic processes. For both martingales and ergodic processes the generalization of the law of large numbers exists in a weak and a strong version. The weak versions, convergence of martingales in mean and the mean ergodic theorem, can be proved using information theoretic methods. The strong versions, point wise convergence of martingales and the individual ergodic theorem, can be proved using very different techniques. In this paper we shall present a unified approach using information divergence and a new maximal inequality.

Let $P$ and $Q$ be probability measures. Then the information divergence from $P$ to $Q$ is defined by

$$D (P \| Q) = \int \log \frac{dP}{dQ} dP \quad \text{if } P \ll Q,$$

This quantity is also called the Kullback-Leibler discrimination or relative entropy. Information divergence does not define a metric, but is related to total variation via Pinsker’s inequality $\frac{1}{2} \|P - Q\|^2 \leq D (P \| Q)$ proved by I. Csiszár [11] and others. If $(P_n)_{n \in \mathbb{N}}$ is a sequence of probability distributions, we say that $(P_n)_{n \in \mathbb{N}}$ converges to $Q$ in information if $D (P_n \| Q) \to 0$ for $n \to \infty$. Pinsker’s inequality shows that convergence in information is a stronger condition than convergence in total variation. See [12] for details about topologies related to information divergence.

It is not a new idea to relate results from probability theory and information theory. Some convergence theorems in probability theory can be reformulated as “the entropy converges to its maximum”. A. Rényi [1] used information divergence to prove convergence of Markov chains to equilibrium on a finite state space. Later I. Csiszár [2] and Kendall [3] extended Rényi’s method to provide convergence on countable state spaces. Their proofs use basically that information divergence is a Csiszár $I$-divergence. Later J. Fritz [4] used information theoretic arguments to establish convergence in total variation of reversible Markov chains. Recently A. Barron [5] improved Fritz’ method and proved convergence in information divergence. Other limit theorems has been proved using information theoretic methods. The Central Limit Theorem was treated by Linnik [6] and A. Barron [7], the Local Central Limit Theorem was treated by S. Takano [8] and Poisson’s law was treated by P. Harremoës [9]. There has also been work strengthening weak or strong convergence to convergence in information divergence. All the above mentioned papers have results of this kind, but also work by A. Barron [5] should be mentioned. Some work has also been done where the limit of a sequence is identified as an information projection. The most important paper in this direction is due to I. Csiszár [10].

Barron used properties of information divergence to prove mean convergence of martingales in [5]. The basic result is that information divergence is continuous under the formation of the intersection of a decreasing sequence of σ-algebras. The same technique can be used to obtain a classical result of Pinsker [13] about continuity under an increasing sequence of σ-algebras. In order to prove almost sure convergence of martingales Barron just remarks that mean convergence implies pointwise convergence according to the maximal inequalities. The main result is the following new maximal inequality.

Theorem 1: Let $X_1, X_2, \ldots, X_n$ be a positive martingale, and put $X^* = \max (X_j)$. If $X_1 = 1$ then

$$E (X^\text{max}) - 1 - \ln E (X^\text{max}) \leq E (X_n \ln (X_n)) .$$

Pointwise convergence of a martingale follows very easily from this kind of maximal inequality if for which $E (X_n \ln (X_n))$ is bounded.

Supported by a Post. Doc. fellowship from the Villum Kann Rasmussen Foundation and by the Danish Natural Science Council.
II. SOME NEW MAXIMAL INEQUALITIES

Let $\Omega$ be a measurable space with a probability measure $P$. All mean values will be calculated with respect to $P$. The following inequalities are well known. Here they are just stated for completeness.

**Lemma 2:** Let $X_1, X_2, ..., X_n$ be a martingale. Put $X^{\max} = \max (X_j)$ and $X^{\min} = \min (X_j)$. Then

$$\lambda \cdot P \left( X^{\max} \geq \lambda \right) \leq E \left( X_n \cdot 1_{X^{\max} \geq \lambda} \right)$$

and

$$\lambda \cdot P \left( X^{\min} \leq \lambda \right) \geq E \left( X_n \cdot 1_{X^{\min} \geq \lambda} \right).$$

**Proof:** Let $\tau$ denote the stopping time $\inf \{k \mid X_k \geq \lambda\}$. Then

$$\lambda \cdot P \left( X^{\max} \geq \lambda \right) = \lambda \cdot P \left( \tau \leq n \right) = E \left( \lambda \cdot 1_{\tau \leq n} \right) \leq E \left( X_\tau \cdot 1_{\tau \leq n} \right) = E \left( X_n \cdot 1_{\tau \leq n} \right).$$

The second inequality is proved in the same way.

**Lemma 3:** Let $X_1, X_2, ..., X_n$ be a positive martingale. Put $X^{\max} = \max (X_j)$ and $X^{\min} = \min (X_j)$. If $X_1 = 1$ then

$$E \left( X^{\max} \right) - 1 \leq E \left( X_n \ln (X^{\max}) \right)$$

and

$$E \left( X^{\min} \right) - 1 \leq E \left( X_n \ln (X^{\min}) \right).$$

**Proof:** By using that $X^{\max} \geq X_1 = 1$ we get

$$E \left( X^{\max} \right) = \int_0^{\infty} P \left( X^{\max} \geq t \right) dt = 1 + \int_1^{\infty} \frac{1}{t} E \left( X_n \cdot 1_{X^{\max} \geq t} \right) dt$$

$$\leq 1 + \int_1^{\infty} \frac{1}{t} E \left( X_n \cdot \int_0^{X^{\max}} 1 \frac{dt}{t} \right)$$

$$= 1 + E \left( X_n \cdot \ln (X^{\max}) \right) dt.$$

Similarly, $0 \leq X^{\min} \leq X_1 = 1$ implies that

$$E \left( X^{\min} \right) = \int_0^1 P \left( X^{\min} \geq t \right) dt = \int_0^1 \left( 1 - P \left( X^{\min} < t \right) \right) dt$$

$$= 1 - \int_0^1 P \left( X^{\min} < t \right) dt \leq 1 - \int_0^1 \frac{1}{t} E \left( X_n \cdot 1_{X^{\min} < t} \right) dt$$

$$= 1 - E \left( X_n \cdot \int_0^{X^{\min}} \frac{1}{t} dt \right) = 1 + E \left( X_n \cdot \ln (X^{\min}) \right) dt.$$

The following function will play an important role in what follows. Put $\gamma (x) = x - 1 - \ln x$ for $x > 0$. Remark that $\gamma$ is strictly convex with a minimum at $x = 1$.

Theorem 1 can now be stated as:

**Theorem 4:** Let $X_1, X_2, ..., X_n$ be a positive martingale. Put $X^{\max} = \max (X_j)$ and $X^{\min} = \min (X_j)$. If $X_1 = 1$ then

$$\gamma \left( E \left( X^{\max} \right) \right) \leq E \left( X_n \ln (X^{\max}) \right)$$

and

$$\gamma \left( E \left( X^{\min} \right) \right) \leq E \left( X_n \ln (X^{\min}) \right).$$

**Proof:** The inequality $\ln (t) \leq \ln (E (X^*)) + t/E (X^{\max}) - 1$ implies that

$$E \left( X^{\max} \right) - 1 \leq E \left( X_n \ln (X^{\max}) \right)$$

$$\leq E \left( X_n \ln (X_n) + E \left( X_n \ln \left( \frac{X^{\max}}{X_n} \right) \right) \right).$$

Inequality (1) is obtained by reorganizing the terms. Inequality (2) is proved in the same way.

**Example 5:** The simplest application of the inequality involves two $\sigma$-algebras and two probability measures $P$ and $Q$. Put $X_1 = 1$ and $X_2 = dP/dQ$. Then $E \left( X^{\max} \right) = 1 + \|P - Q\|/2$ and $E \left( X^{\min} \right) = 1 - \|P - Q\|/2$. Inequality 1 states that

$$\frac{1}{2} \|P - Q\| - \ln \left( 1 + \frac{1}{2} \|P - Q\| \right) \leq D (P \| Q).$$

This inequality was proved by Volkonskij and Rozanov [15] and can be considered as a weak version of Pinsker’s inequality, see [16] for more details about the history of this problem.
which gives a more tight bound on total variation than Inequality (3). If $D(P||Q) \geq 2$ then Pinsker’s Inequality gives no restriction on the total variation, but Inequality (4) will give a bound even for large values of $D(P||Q)$.

The results have been stated for martingale $X_n$, where $n$ belong to a finite set. It is straightforward to generalize the inequalities to larger sets.

We shall see that Inequality (2) is tight. Thus we have to consider a martingale for which the points on the curve $y = \gamma(x)$ are attained. To construct such an example the time has to be continuous. Let the unit interval $[0; 1]$ be equipped with the Lebesgue measure. Let $\mathcal{F}_t$ be the $\sigma$-algebra generated by the set $[0; t]$ and the set Borel measurable subsets of $[t; 1]$. Let $f$ denote the function (random variable) on $[0; 1]$ given by

$$f(x) = (\beta + 1)x^\beta$$

where $\beta > 0$. Remark that $f$ is increasing. The conditional expectation of $f$ given $\mathcal{F}_t$ is

$$\mathbb{E}(f | \mathcal{F}_t)(x) = \begin{cases} \frac{\int_0^t (x-y) \, dy}{f(x)} & \text{for } x < t \\ f(x) & \text{for } x \geq t \end{cases}$$

If $f_t(x) = \mathbb{E}(f | \mathcal{F}_t)(x)$ then $f_t$ is a reversed martingale in $t$.

Put

$$f_{\min}^m = \inf_{t \in [0; 1]} f_t(x).$$

The infimum is attained for $t = x$. Thus

$$f_{\min}^m(x) = \int_0^x \frac{f(y) \, dy}{y} = \int_0^x \frac{(\beta + 1)y^\beta \, dy}{y} = x^\beta = \frac{f(x)}{\beta + 1}.$$ 

Thus

$$E(f_{\min}^m) = E\left(\frac{f(x)}{\beta + 1}\right) = \frac{1}{\beta + 1}.$$ 

Hence $\beta = \frac{1}{E(f_{\min}^m)} - 1$, and this implies that

$$\int_0^1 f(x) \ln(f(x)) \, dx = \ln(\beta + 1) - \frac{\beta}{\beta + 1} = \gamma(E(f_{\min}^m)).$$

If $\beta$ in the above example is element in $[-1; 0]$ then $f$ is decreasing and $f_{\min}^m$ shall be replaced by $f_{\max}^m$. All the calculations are the same, and therefore also Inequality (1) is tight.

### III. Convergence of Martingales

In order to prove convergence of martingales we have to reorganize our inequalities somewhat.

#### Theorem 6: Let $X_1, X_2, \ldots$ be a martingale and assume that $E(X_j) = 1$. Let $Q_j$ be the probability measure given by $\frac{dQ_j}{dP} = X_j$. For $m \leq n$ put $X_{m,n}^{\max} = \sup_{j=m,\ldots,n} X_j$ and put $\gamma(x) = x - 1 - \ln(x)$. Then

$$\gamma(E(X_{m,n}^{\max})) \leq D(Q_n||Q_m).$$

**Proof:** For each value $x$ of $X_n$ we have

$$\gamma\left(E\left(\frac{X_{m,n}^{\max}}{X_m} | X_m = x\right)\right) \leq E\left(\frac{X_n}{X_m} \ln\frac{X_n}{X_m} | X_m = x\right) = E\left(X_n \ln\frac{X_n}{X_m} | X_m = x\right) = D(Q_n||Q_m).$$

Again a similar inequality is satisfied for the minimum of a martingale, i.e.

$$\gamma(E(X_{m,n}^{\min})) \leq D(Q_n||Q_m).$$

Now, let $X_1, X_2, \ldots$ be a martingale. Without loss of generality we will assume that $E(X_n) = 1$. Then

$$E(X_n \ln X_n) - E(X_n \ln X_m) = D(Q_n||Q_m).$$

We see that $E(X_n \ln X_n)$ is increasing. Assume that $E(X_n \ln X_n)$ is bounded. Then $D(Q_n||Q_m)$ converges to $0$ for $m, n$ tending to infinity. In particular $E(X_{m,n}^{\max} - X_{m,n}^{\min}) \rightarrow 0$ for $m, n \rightarrow \infty$. Thus, $P(X_{m,n}^{\max} - X_{m,n}^{\min} \geq \varepsilon) \rightarrow 0$ for $m, n \rightarrow \infty$ and $X_n$ is a Cauchy sequence with probability one. Therefore the martingale converges point wise almost surely.

### IV. The Individual Ergodic Theorem

The maximal inequalities for martingales played a central role in the proof of Inequality (5). We shall use the Maximal Ergodic Theorem to prove a similar inequality for ergodic sequences. First we shall state the Maximal Ergodic theorem.

Let $(\Omega, \mathcal{B}, Q)$ be a probability space with a measure preserving transformation $T : \Omega \rightarrow \Omega$. Let $f$ be a random variable with $E(|f| < \infty$. Put

$$f_{\max}^m = \max_{1 \leq k \leq n} \frac{1}{k} \sum_{j=0}^{k-1} f \circ T^j$$

and

$$f_{\max}^\infty = \sup_n f_{\max}^n.$$

**Theorem 7—The Maximal Ergodic Theorem:** Let $g$ be an invariant function on $\Omega$ with $E(|g| < \infty$. Then

$$E(\max_g f \cdot 1_{f_{\max}^\infty > g}) \geq 0.$$
The Individual Ergodic Theorem can be formulated as follows.

**Corollary 8—The Individual Ergodic Theorem:** The sequence \( \frac{1}{n} \sum_{j=1}^{k} f \circ T^j \) converges almost surely.

We shall now give quantitative bounds for the rate of convergence of log-bounded random variables. Let \( T \) be an ergodic transformation. Assume that \( f \) is a non-negative random variable. We shall assume that \( E(f) = 1 \) and that \( E(f \log(f)) < \infty \). A sequence of probability measures \( P_n \) are defined by

\[
dP_n \over dQ = \frac{1}{k} \sum_{j=0}^{k-1} f \circ T^j.
\]

We have \( D(P_n || Q) = E(f \log(f)) \) and therefore \( D(P_n || Q) \to 0 \). Now using that we have a maximal inequality we are able to repeat the proofs of the bounds for martingales.

\[
S_n^{\text{max}} = \sup_{k \geq n} \frac{1}{k} \sum_{j=0}^{k-1} f \circ T^j
\]

\[
S_n^{\text{min}} = \inf_{k \geq n} \frac{1}{k} \sum_{j=0}^{k-1} f \circ T^j.
\]

This leads us to the inequalities

\[
E(g(S_n^{\text{max}})) \leq D(P_n || Q)
\]

\[
E(g(S_n^{\text{min}})) \leq D(P_n || Q).
\]

Thus convergence of \( P_n \) in information implies almost sure point wise convergence of the densities \( dP_n \over dQ \). One easily proves that \( D(P_n || Q) \to 0 \) for \( n \to \infty \), if and only if \( D(P_n || Q) \to 0 \).

**V. Discussion**

Theorem 1 can be seen as a strengthening of a classical maximal inequality by Doob, which states that

\[
E(X^{\text{max}}) \leq \frac{e}{e-1} (1 + E(X_n \ln(X_n))).
\]

A plot comparing the two inequalities displays that Doob’s inequality corresponds to a tangent to the function \( g \). Thus the new inequality is superior to Doob’s inequality in a neighborhood of 1, and it the behavior in this region that implies convergence of the martingale. Only in a neighborhood of \( E(X^{\text{max}}) = e \) Doob’s inequality is optimal.

Martingale theory and ergodic theory generalizes the law of large numbers, so one may be interested in applications of the new inequalities in the law of large numbers. Let \( X_1, X_2, \ldots \) be a sequence of independent identically distributed positive random variables such that \( E(X_1 \ln(X_1)) < \infty \). Then \( (X_1 + X_2 + \ldots + X_n) / n \) is a reversed martingale and the new inequality states that

\[
\gamma \left( E \sup_{n > N} \frac{X_1 + X_2 + \ldots + X_n}{n} \right) \leq E \left( \frac{X_1 + X_2 + \ldots + X_N}{N} \ln \left( \frac{X_1 + X_2 + \ldots + X_N}{N} \right) \right) \leq E \left( X_1 \ln \left( 1 + \frac{X_1 - 1}{N} \right) \right).
\]

If we use that the sequence \( X_1, X_2, \ldots \) is ergodic then we arrive to exactly the same inequality.

In this paper upper bounds for \( E(X^{\text{max}}) \) and lower bounds on \( E(X^{\text{min}}) \) are given in terms of \( E(X_n \ln(X_n)) \), and each of the bounds is shown to be tight. In the example, the tightness is obtained for a certain value of a parameter \( \beta \). The upper and the lower bounds are tight for different values of \( \beta \). Therefore a tighter bound on \( E(X^{\text{max}} - X^{\text{min}}) \) is possible in terms of \( E(X_n \ln(X_n)) \). Such tighter bounds would be highly interesting and an obvious subject for further investigation. Convergence of martingales and the Individual Ergodic Theorem are used in the Sandwich-proof of the Shannon-McMillan-Breiman theorem [14]. One should expect that the bounds presented in this paper will lead explicit bounds in this theorem.

One should also remark that convergence of Markov chains could be proved using ideas from the theory of information projections as pointed out by Barron in [5]. Let \( X_1, X_2, \ldots \) be a homogeneous Markov chain with Markov kernel \( M \) and stationary distribution \( Q \). Let \( P \) denote the initial distribution of \( X_1 \). Then our inequality states that

\[
\gamma \left( \sup_{n \geq N} \frac{dM^n P}{dQ} \right) \leq D \left( M^n P || Q \right).
\]

Thus convergence in information implies point wise convergence of the densities.

In this manuscript the focus has been on random variables which forms martingales. We have seen that the upper bound of \( g(E(X^{\text{max}})) \) can be given in terms of the information divergence from one probability measure to another, but also the quantity \( E(X^{\text{max}}) \) can be given in terms of measures. In general the random variable \( X^{\text{max}} \) is not a probability density, but
it defines a measure \(\frac{\bigvee_{i=1}^{n} Q_i}{dQ} = X^{\text{max}},\)

and

\[E(\chi^{\text{max}}) = \left(\bigvee_{i=1}^{n} Q_i\right)(\Omega).\]

Hence we have

\[g\left(\left(\bigvee_{i=1}^{n} Q_i\right)(\Omega)\right) \leq D(Q_n||Q_1).\]  

One may ask if this inequality also holds in other situations than martingales and ergodic sequences.

For \(n = 2\) Inequality (6) always holds as we have seen in Example 5. For \(n \geq 3\) Inequality (6) does not hold in general.

Let \(X_1, X_2, \ldots\) be a martingale with respect to \(\mathbb{F}_1, \mathbb{F}_2, \ldots\). Assume that \(Y_k\) is \(\mathbb{F}_k\) measurable. Then for \(l \geq k\)

\[E_{Q_{k}}(Y_k) = E_{P_{k}}\left(Y_k \cdot \frac{dQ_{k}}{dP}\right) = E_{P_{k}}(Y_k \cdot X_k) = E(Y_k \cdot E(X_k | X_k)) = E(Y_k \cdot X_l) = E_{Q_l}(Y_k).\]

Thus \(Q_k\) and \(Q_l\) have the same restriction to \(\mathbb{F}_k\). Actually \(Q_k\) is the information projection of \(Q_l\) into the convex set of probability measures \(Q\) such that \(Q\) and \(Q_l\) have the same restriction to \(\mathbb{F}_k\). With these observations we are able to state the following conjecture.

**Conjecture 9:** Let \(P\) be a probability measure and let \(C_1, C_2, \ldots, C_n\) be a decreasing sequence of convex sets. Let \(Q_k\) denote the information projection of \(P\) on \(C_k\). Then

\[g\left(\left(\bigvee_{i=1}^{n} Q_i\right)(\Omega)\right) \leq D(Q_n||Q_1).\]  

Similar conjectures can be formulated for a sequence of maximum likelihood estimates and for the minimum of measures instead of the maximum. The conjecture is true for \(n = 2\) and in a number of other special cases. If the conjecture is true a number of new convergence theorems will hold. For instance a sequence of maximum entropy distributions on smaller and smaller sets will converge point wise almost surely if some weak regularity conditions are fulfilled.

**REFERENCES**


