Martingales and Information Divergence

Peter Harremoës

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 convergence of $X \log X$ bounded martingales. Application in to the Shannon-McMillan-Breiman theorem and Markov chains are mentioned.

Index Terms—Information divergence, martingale, convergence, maximal inequality.

I. INTRODUCTION

Comparing results from probability theory and information theory is not a new idea. 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 an $I$-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 in total variation for 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 a information projection. The most important paper in this direction is due to I. Csiszár [10].

The law of large numbers exists in different versions. The weak law of large numbers states that an empirical average converges to the true 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, pointwise convergence of martingales and the individual ergodic theorem, can be proved using very different techniques. In this paper we shall see that pointwise convergence of martingales can also be proved using information theoretic methods.

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

$$D(P \parallel Q) = \begin{cases} \int \log \frac{dP}{dQ} \, dP & \text{if } P \ll Q, \\ \infty & \text{otherwise.} \end{cases}$$

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. Other limit theorems have 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 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].

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. Using 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

University of Copenhagen.
Lemma 3: Let \( X_1, X_2, \ldots, X_n \) be a positive martingale. Put \( X_{\text{max}} = \max(X_j) \) and \( X_{\text{min}} = \min(X_j) \). If \( X_1 = 1 \) then

\[
E(X_{\text{max}}) - 1 \leq E(X_n \ln(X_{\text{max}}))
\]

(12)

and

\[
E(X_{\text{min}}) - 1 \leq E(X_n \ln(X_{\text{min}})) .
\]

(13)

Proof: By using that \( X_{\text{max}} \geq X_1 = 1 \) we get

\[
E(X_{\text{max}}) = \int_0^\infty P(X_{\text{max}} \geq t) \, dt
\]

(14)

\[
= 1 + \int_1^\infty P(X_{\text{max}} \geq t) \, dt
\]

\[
\leq 1 + \int_1^\infty \frac{1}{t} E(X_n \cdot 1_{X_{\text{max}} \geq t}) \, dt
\]

(15)

\[
= 1 + E \left( \int_1^\infty \frac{X_n \cdot 1_{X_{\text{max}} \geq t}}{t} \, dt \right)
\]

(16)

\[
= 1 + E(X_n \cdot \ln(X_{\text{max}})) \, dt).
\]

Similarly, \( 0 \leq X_{\text{min}} \leq X_1 = 1 \) implies that

\[
E(X_{\text{min}}) = \int_0^1 P(X_{\text{min}} \geq t) \, dt
\]

(17)

\[
= \int_0^1 (1 - P(X_{\text{min}} < t)) \, dt
\]

(18)

\[
= 1 - \int_0^1 P(X_{\text{min}} < t) \, dt
\]

(19)

\[
\leq 1 - \int_0^1 \frac{1}{t} E(X_n \cdot 1_{X_{\text{min}} < t}) \, dt
\]

(20)

\[
= 1 - E \left( \int_0^1 \frac{X_n \cdot 1_{X_{\text{min}} < t}}{t} \, dt \right)
\]

(21)

\[
= 1 - E(X_n \cdot \frac{1}{X_{\text{min}}} \, dt)
\]

(22)

\[
= 1 + E(X_n \cdot \ln(X_{\text{min}})) \, dt).
\]

Inequality (27) is obtained by reorganizing the terms. Inequality (28) is proved in the same way.

Now we shall see that Inequality (27) is tight. Thus we have to consider a martingale for which the points on the curve \( y = g(x) \) 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 Borel measurable sets on \([0; t]\) and the set \([t; 1]\). Let \( f \) denote the function (random variable) on \([0; 1]\) given by

\[
f(x) = (\beta + 1) (1 - x)^\beta
\]

(34)
where $\beta \in [-1; 0]$. Remark that $f$ is increasing. The conditional expectation of $f$ given $\mathbb{F}_t$ is

$$
\mathbb{E}(f \mid \mathbb{F}_t)(x) = \begin{cases} 
\frac{f(x)}{\ln f(x)} & \text{for } x < t \\
\frac{f'(x)}{1-t} & \text{for } x \geq t.
\end{cases}
$$

(35)

Thus $\beta = \frac{1}{\ln(f(x))} - 1$. This implies that

$$
\int_0^1 f(x) \ln(f(x)) \, dx = \int_0^1 (\beta + 1) (1-x)^\beta \ln ((\beta + 1) (1-x)^\beta) \, dx
$$

$$
= \ln(\beta + 1) - \frac{\beta}{\beta + 1} = g(E(f^{\max}).
$$

(43)

If $\beta$ in the above example is positive then $f$ is decreasing and $f^{\max}$ shall be replaced by $f^{\min}$. All the calculations are the same, and therefore also Inequality (28) is tight.

III. CONVERGENCE OF MARTINGALES

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

**Theorem 5:** 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 $\frac{dQ_j}{df} = X_j$. For $m \leq n$ put $X^{\max}_{m,n} = \sup_{j=m,\ldots,n} X_j$ and put $g(x) = x - 1 - \ln(x)$. Then

$$
g(E(X^{\max}_{m,n})) \leq D(Q_n||Q_m).
$$

(44)

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

$$
g(E\left(\frac{X^*}{X_m} \mid X_m = x\right)) \leq E\left(\frac{X_n}{X_m} \ln\left(\frac{X_n}{X_m}\right) \mid X_m = x\right).
$$

(45)

Using convexity of $g$ leads to

$$
E\left(X_m g\left(E\left(\frac{X^*}{X_m} \mid X_m\right)\right)\right)
\leq E\left(X_m E\left(\frac{X_n}{X_m} \ln\left(\frac{X_n}{X_m}\right) \mid X_m\right)\right)
= E\left(X_m \ln\left(\frac{X_n}{X_m}\right)\right)
= D(Q_n||Q_m).
$$

(46)

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

$$
g(E(X^{\min}_{m,n})) \leq D(Q_n||Q_m).
$$

(47)

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

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

(48)

We see that $E(X_n \ln X_n)$ is increasing. Assume that $E(X_n \ln X_n)$ is not bounded. Then $D(Q_n||Q_m)$ converges to $0$ for $m, n$ tending to infinity. In particular $E(X^{\max}_{m,n} - X^{\min}_{m,n}) \to 0$ for $m, n \to \infty$. Thus, $P(X^{\max}_{m,n} - X^{\min}_{m,n} \geq \varepsilon) \to 0$ for $m, n \to \infty$ and $X_n$ is a Cauchy sequence with probability one. Therefore the martingale converges pointwise almost surely.
defines a measure \( \bigvee_{i=1}^{n} Q_i \) given by
\[
\frac{d}{dQ} \left( \bigvee_{i=1}^{n} Q_i \right) = X^{\text{max}},
\]
and
\[
E \left( X^{\text{max}} \right) = \left( \bigvee_{i=1}^{n} Q_i \right) (\Omega).
\]
Hence we have
\[
g \left( \left( \bigvee_{i=1}^{n} Q_i \right) (\Omega) \right) \leq D \left( Q_n || Q_1 \right).
\]
One may ask if this inequality also holds in other situations. If \( n = 2 \) we have
\[
(Q_1 \vee Q_2) (\Omega) = 1 + \frac{1}{2} ||Q_1 - Q_2||,
\]
where the norm is the total variation norm. Then Inequality 52 states that
\[
\ln \left( 1 + \frac{1}{2} ||Q_1 - Q_2|| \right) - \frac{1}{2} ||Q_1 - Q_2|| \leq D \left( Q_2 || Q_1 \right).
\]
This inequality was proved by Volkonskij and Rozanov [15] and was later refined to Pinsker’s inequality, see [16] for more details about the history of this problem.

For \( n \geq 3 \) Inequality 52 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 \( l \geq k \)
\[
E_{Q_k} (Y_k) = E_{P} \left( Y_k \cdot \frac{dQ_k}{dP} \right)
\]
\[
= E_{P} (Y_k \cdot X_k)
\]
\[
= E \left( Y_k \cdot E \left( X_k \mid X_l \right) \right)
\]
\[
= E \left( E \left( Y_k \cdot X_l \mid X_k \right) \right)
\]
\[
= E \left( Y_k \cdot X_l \right)
\]
\[
= EQ_k \left( Y_k \right).
\]
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 6:** 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
\[
g \left( \left( \bigvee_{i=1}^{n} Q_i \right) (\Omega) \right) \leq D \left( Q_n || Q_1 \right).
\]

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 maximum entropy distributions on smaller and smaller sets will converge pointwise almost surely if some weak regularity conditions are fulfilled.
REFERENCES


