An application of martingales in the limit to a problem in information science

AN APPLICATION OF MARTINGALES IN THE LIMIT TO A PROBLEM IN INFORMATION SCIENCE

by

L. Egghe

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]

and

UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium

ABSTRACT

Martingales in the limit (mils) were introduced about two decades ago as non-trivial extensions of martingales. It was proved in 1976 that they have good convergence properties (at least) for real-valued stochastic processes. But so far there have not been found any "real-life" applications of mils.

In this article we apply the full generality of mils to a problem in information science. There we study the evolution in time of source journals as e.g. defined by the Institute for Scientific Information (ISI) who selects, on a yearly basis, the most "visible" journals in the world. In this connection one also encounters quasi martingales.

martingale in the limit, mil, quasi martingale, information science, source journal

AMS 1991 Subject Classification :

Primary : 60B12

Secondary : 94A15

[1]Permanent address

0. Introduction

About 25 years ago some probabilists introduced generalisations of martingales and studied their convergence properties, both in $\mathbb{R}$ as in general infinite dimensional Banach spaces. We can refer to [1] and [2] for two original references and to the books [5] and [4] of Egghe resp. Edgar and Sucheston which are entirely devoted to these extensions.

The main idea behind these extensions is to prove convergence of a stochastic proces by looking at the behaviour of the conditional expectations. Let us first look at the simplest case : the case of a martingale. First we fix some notation. Let $(\Omega,\mathcal{F},P)$ be a probability space. Let L^1 denote the space $L^1(\Omega,\mathcal{F},P)$ of all integrable real valued functions on Ω. If $f\in L^1$ and $G\subset\mathcal{F}$ is a sub-σ-algebra of the σ-algebra $\mathcal{F}$, then $E(f|G)$ or $E^{G}f$ denotes the conditional expectation of f with respect to G.

Consider a sequence $(X_n,\mathcal{F}_n)_{n\in\mathbb{N}}$ where $X_n\in L^1$, $\forall n\in\mathbb{N}$ and $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is an increasing sequence of sub-σ-algebras of $\mathcal{F}$. We say that $(X_n,\mathcal{F}_n)_{n\in\mathbb{N}}$ is a stochastic process (or adapted sequence) if every X_n is $\mathcal{F}_n$-measurable.

If $E^{\mathcal{F}_n}X_{n+1}=X_n$, a.e. for every $n\in\mathbb{N}$, we say that the process is a martingale. If the equality sign is replaced by $\geq$ we call the process a submartingale and if we replace it by $\leq$ we call it a supermartingale. These processes are well-known in probability theory (see [12]) and their applications are numerous, also beyond the mathematical scene. In fact, we can mention here that we were able to apply (sub-) (super-) martingale theory to the evolution (e.g. growth) of databases in information science. See for this [6], [7], [9] and [10].

The processes described above all have the property to have conditional expectations that behave in a "monotonic" way. That lead several probabilists to the definition of processes for which the difference

$$E^{\mathcal{F}_n}X_{n+1} - X_n \qquad (1)$$

(or using other indices - see below) does not not have a fixed sign. Instead one requires that (1) goes to zero (in a certain way) if the index goes to infinity.

We encounter the following extensions of martingales (see [4] or [5]),

1. The martingale in the limit (mil)

   A stochastic process $(X_n, \mathcal{F}_n, P)$ is called a martingale in the limit (mil, for short) if

$$\lim_{m\to\infty} \sup_{\substack{n\geq m \\ n\in\mathbb{N}}} |E^{\mathcal{F}_m} X_n - X_m| = 0, \text{ a.e.} \tag{2}$$

   This clearly generalises martingales since for them $E^{\mathcal{F}_m} X_n = X_m$, a.e. for all $n,m\in\mathbb{N}$, $m\leq n$. We can mention here the theorem of Mucci of 1976, [11], starting that L^1-bounded mils converge a.e..

2. Asymptotic martingales (amarts) and uniform amarts

   Let $(X_n, \mathcal{F}_n, P)$ be a stochastic process and denote by T the set of all bounded $\mathbb{N}$-valued stopping times, i.e. functions σ for which $\{\sigma=n\}\in\mathcal{F}_n$, $\forall n\in\mathbb{N}$. Let $\sigma\in T$. Denote by X_σ the function

$$(X_\sigma)(\omega) = X_{\sigma(\omega)}(\omega) \tag{3}$$

   for $\omega\in\Omega$. We say that $(X_n, \mathcal{F}_n, P)$ is an asymptotic martingale (amart, shortly) if the directed net

$$\left(\int_\Omega X_\sigma \right)_{\sigma\in T} \tag{4}$$

   converges (using the natural order on T). Note that also amarts generalise martingales : for them the net (4) is constant. Condition (4), and for real-valued processes (as we assume here) the notion of amart is equivalent to the one of uniform amart : $(X_n, \mathcal{F}_n, P)$ is called a uniform amart if

$$\lim_{\sigma\in T} \sup_{\substack{\tau\geq\sigma \\ \tau\in T}} E(|E^{\mathcal{F}_\sigma}X_\tau - X_\sigma|)=0. \tag{5}$$

For the proof of this we refer the reader to [4] or [5]. The equivalence of amarts and uniform amarts in fact characterises finite dimensional Banach spaces as can be read in [5]. Also in [5] one shows that quasi martingales are uniform amarts. A quasi martingale $(X_n, \mathcal{F}_n, P)$ *is a stochastic process for which*

$$\sum_{n=1}^{\infty} E(|E^{\mathcal{F}_n}X_{n+1} - X_n|<\infty. \tag{6}$$

They converge since we have more generally, by a theorem of Bellow from 1978 [2], that any L^1-bounded uniform amart converges a.e..

As mentioned to me by G. Edgar ([3]) there have not been found any real-life applications of mils or amarts in the sense that we are in need of their full generality.

In this paper we will construct a stochactic process that describes the evolution of a set of source journals, e.g. a set of internationally visible ("important") journals as they are defined on a yearly basis by the company ISI (Institute for Scientific Information). This will be explained in the next section.

The paper closes with some problems in this connection.

1. The evolution of a set of source journals

The Institute for Scientific Information (ISI) in Philadelphia determines, on a yearly-basis, its set of so-called "source journals", i.e. the set of, according to their standards, most visible journals in the world. The exact criterium on which one decides whether or not a journal becomes a source journal is of no importance here. Let us just mention that their decisions are

based on citation analysis, within each subject. We stress the fact that our model, to be developed here, allows for any criterium.

Such a list of source journals then forms the basis for many evaluation studies of scientific research. Although widely accepted (especially in the exact sciences, applied sciences and medical sciences) and applied (e.g. in the allocation of research budgets), based on the degree of visibility of the journal in which one publishes, there is also a lot of criticism on the method (see e.g. [8]). Let us mention one problem. One often argues that the list of source journals is not complete because several important journals are not included. This criticism is often heard also in developing countries where they claim to have valuable journals that are not recognised as source journals. This might be true in some cases and false in others. One could of course argue that, in an ideal world, even if one starts with a partially "wrong" set of source journals, we would end up, eventually, with "the right set". This is so because journals that are little visible will cite the most visible journals and because the degree of citedness is the basis of the selection as source journal, those visible journals that were not in the initial list will be included sooner or later.

Many problems can be posed here. For a list of them we refer the reader to [13] and to the last section. Here we will limit our study to the following problem. Suppose, for a start, that we have a universe U of "all" journals. Here we can include all journals that exist (or came into existence) in the time period under study. Of course, in this setting, a journal cannot become a source journal before it actually exists.

Suppose, at the starting point we select a certain subset A of U. How this set is determined is of no importance but its "survival", when time passes, will be determinded by the criteria that are adopted in the decision for a source journal (e.g. citation criteria). As explained above this set A will change, when time passes due to the evolution in the visibility of journals in U and the possible fact that in A some "local" journals could have been wrongly selected (dependent on where the initial set is defined).

Problem : When time passes, what will be the remainder of A? In other words, what will be left over from our initial set of source journals?

Note that this process is not necessarily increasing or decreasing. Indeed, when going from the first year to the second one there is a decrease : some journals of A can be dropped as source journals . But already from year 2, and continuing so, journals from A can be added or deleted (the ones added in case they were previously deleted from A at least once).

So, the evolution of the "remainder" of A, the problem studied here, does not seem to be a process where expectations in the future are inferior or superior to what we have at a certain time (as is the case for (super-) (sub-) martingales). We will now construct the stochastic process that describes the remainder of A.

Let us hence take $A \subset U$ at time t=1. If we take the unity of time to be a year, say, then we have that several journals in A might leave as a source journal and also that many other journals from U become a source journal. This is not easy to model. Therefore we assume that the step of increase by one (from t to t+1) stands for <u>one</u> change in the set of source journals (of any type : one in or one out). This is not a restriction since the many transactions in a year can be subdivided as indicated above.

We define

X_t = the number of journals from A that, after t steps, are source journals.

Note that X_t represents a snapshot at time t (i.e. after t steps). The number X_t refers to journals of A that stayed as source journal all the time as well as to journals of A that left as a source journal but then (before or at t) were again picked up as a source journal. Note that $X_1 = \#A$.

At each t we denote by A_t the set of source journals. Hence $X_t = \#(A_t \cap A)$, $\forall t \in \mathbb{N}$. To go from t to t+1 we have the following algorithm :

- with probability $\alpha(t)$ there will be a journal from $A_t \cap A$ that leaves as a source journal at t+1.

If this is not the case (probability $1-\alpha(t)$) there are two possibilities :

- with probability $\beta(t)$ a journal belonging to $A \backslash A_t$ re-enters the set of source journals at $t+1$

- with probability $1-\beta(t)$ it will be a journal from $U \backslash A$ that enters as a source journal at $t+1$.

This description also determines the stochastic process $(X_t, \mathcal{F}_t)_{t \in \mathbb{N}}$ on the probability space $(\Omega, \mathcal{F}, P)$. We have the following equation.

$$E^{\mathcal{F}_t} X_{t+1} = \alpha(t)(X_t - 1) + (1 - \alpha(t))[\beta(t)(X_t + 1) + (1 - \beta(t))X_t].$$

Hence

$$E^{\mathcal{F}_t} X_{t+1} = X_t + \beta(t) - \alpha(t) - \alpha(t)\beta(t). \tag{7}$$

Note that $\beta(1)=0$. In general, $\alpha(t)$, $\beta(t)$ are r.v.s on $(\Omega, \mathcal{F}, P)$. It is clear that equation (7) determines a process that allows for $E^{\mathcal{F}_t} X_{t+1} \geq X_t$ as well as $E^{\mathcal{F}_t} X_{t+1} \leq X_t$. The only formal limitation on the $\alpha(t)$s and $\beta(t)$s is that, $\forall t \in \mathbb{N}$

$$-1 \leq \beta(t) - \alpha(t) - \alpha(t)\beta(t) \leq 1. \tag{8}$$

If one applies this to its maximal possibility and from t=1 on we even have that, at certain times, we have the whole of A as source journal and at other times we have nothing of A left as source journals. Hence we have here a divergent process.

But even more "moderate" behaviour of the $\alpha(t)$s and $\beta(t)$s does not always lead to a convergent process. If we, e.g., require that

$$\lim_{t \to \infty} (\beta(t) - \alpha(t) - \alpha(t)\beta(t)) = 0, \text{ a.e.} \tag{9}$$

we are dealing with a process with the property

$$\lim_{t\to\infty}|E^{\mathcal{F}_t}X_{t+1}-X_t|=0,\ \text{a.e.} \tag{10}$$

and it is well-known that examples of such processes exist for which $(X_t)_{t\in\mathbb{N}}$ does not converge (cf. [4] and [5]).

We will determine two cases in which convergence is obtained. It will turn out that we encounter the full generality of mils and quasi martingales.

1.2 $(X_t, \mathcal{F}_t)_{t\in\mathbb{N}}$ as a mil.

Theorem : If

$$\sum_{t=1}^{\infty} \alpha(t) < \infty,\ \text{a.e.} \tag{11}$$

$$\sum_{t=1}^{\infty} \beta(t) < \infty,\ \text{a.e.} \tag{12}$$

then we have that $(X_t, \mathcal{F}_t)_{t\in\mathbb{N}}$ is a mil converging a.e. to an integrable function.

Proof : For every t, t'$\in\mathbb{N}$, t'$\geq$t :

$$E^{\mathcal{F}_t}X_{t'}-X_t=E^{\mathcal{F}_t}(\sum_{i=t}^{t'-1}(\beta(i)-\alpha(i)-\alpha(i)\beta(i))) \tag{13}$$

as follows readily from (7) by induction. Now

$$E^{\mathcal{F}_t}(\sum_{i=t}^{t'-1}\beta(i)) \leq E^{\mathcal{F}_t}(\sum_{i=t}^{\infty}\beta(i))$$

$$\leq E^{\mathcal{F}_t}(\sum_{i=1}^{\infty}\beta(i))-E^{\mathcal{F}_t}(\sum_{i=1}^{t-1}\beta(i))$$

$$= E^{\mathcal{F}_t}(\sum_{i=1}^{\infty}\beta(i))-\sum_{i=1}^{t-1}\beta(i) \tag{14}$$

since $\beta(i)$ is $\mathcal{F}_i$-measurable, by the inductive construction of the process $(X_t,\mathcal{F}_t)_{t\in\mathbb{N}}$.
Now $(E^{\mathcal{F}_t}(\sum_{i=1}^{\infty}\beta(i)),\mathcal{F}_t)$ is an elementary martingale which converges to $\sum_{i=1}^{\infty}\beta(i)$, a.e. cf. [5], theorem II.1.6, p. 24. Hence (14) implies that

$$\lim_{t\to\infty}\sup_{\substack{t'\geq t\\ t'\in\mathbb{N}}} E^{\mathcal{F}_t}(\sum_{i=t}^{t'-1}\beta(i))=0,\ \text{a.e.} \tag{15}$$

The same argument yields

$$\lim_{t\to\infty}\sup_{\substack{t'\geq t\\ t'\in\mathbb{N}}} E^{\mathcal{F}_t}(\sum_{i=t}^{t'-1}\alpha(i))=0,\ \text{a.e.} \tag{16}$$

(11) and (12) imply that also

$$\sum_{t=1}^{\infty}\alpha(t)\beta(t) < \infty,\ \text{a.e.} \tag{17}$$

by the comparison test for the convergence of series and since all $\alpha(t)$, $\beta(t)$ are positive and inferior to 1. Hence the same argument yields

$$\lim_{t\to\infty}\sup_{\substack{t'\geq t\\ t'\in\mathbb{N}}} E^{\mathcal{F}_t}(\sum_{i=t}^{t'-1}\alpha(i)\beta(i))=0,\ \text{a.e.} \tag{18}$$

In conclusion, by (13),

$$\lim_{t\to\infty}\sup_{\substack{t'\geq t\\ t'\in\mathbb{N}}} |E^{\mathcal{F}_t}X_{t'}-X_t|=0,\ \text{a.e.} \tag{19}$$

hence $(X_t,\mathcal{F}_t)$ is a mil. It is uniformly bounded (by #A) and hence the theorem of Mucci applies [11], showing that $(X_t)_{t\in\mathbb{N}}$ converges a.e. to an integrable function. □

1.3 $(X_t,\mathcal{F}_t)_{t\in\mathbb{N}}$ as a quasi martingale.

Theorem : If

$$\sum_{t=1}^{\infty} E(\alpha(t)) < \infty \tag{20}$$

$$\sum_{t=1}^{\infty} E(\beta(t)) < \infty \tag{21}$$

then we have that $(X_t,\mathcal{F}_t)_{t\in\mathbb{N}}$ is a quasi martingale converging a.e. to an integrable function.

Proof : the proof is easy since, $\forall t\in\mathbb{N}$

$$E(|E^{\mathcal{F}_t}X_{t+1}-X_t|)$$

$$= E(|\beta(t)-\alpha(t)-\alpha(t)\beta(t)|)$$

$$\leq E(\beta(t))+E(\alpha(t))+E(\alpha(t)\beta(t)).$$

Hence (20), (21) and again by the comparison test for series we have that

$$\sum_{t=1}^{\infty} E(|E^{\mathcal{F}_t}X_{t+1}-X_t|) < \infty$$

Hence $(X_t, \mathcal{F}_t)_{t\in\mathbb{N}}$ is a uniformly bounded quasi martingale which hence converges (use e.g. Bellow's theorem on the convergence of L^1-bounded uniform amarts [2]) to an integrable function. □

Note : From the above proof it is clear that it is sufficient to require

$$\sum_{t=1}^{\infty} E(|\beta(t)-\alpha(t)-\alpha(t)\beta(t)|) < \infty.$$

2. Problems and suggestions for further research.

2.1 Studying the evolution of $A \subset U$ as the set of source journals requires the study of the process

$$X_t = f(t)$$

where f(t) is a distribution on $(\Omega, \mathcal{F}, P)$. Hence we are dealing here with stochastic processes in possibly infinite dimensional Banach spaces, see [4], [5].

2.2 In the above sense, will we be able to prove that, no matter with which set A we start, we will always end up with the same limit set (i.e. fixed limit distribution)?

2.3 What is the stability of the results obtained here, i.e. if we change the criteria to become a source journal a little bit, will the set of source journals, when followed for t increasing, experience dramatic changes or not?

2.4 Find a condition on $(X_t, \mathcal{F}_t)$ as in (7) such that it becomes an amart (and not necessarily a quasi martingale).

References

[1] D.G. Austin, G.A. Edgar and A. Ionescu Tulcea. Pointwise convergence in terms of expectations. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 30, 17-26, 1974.

[2] A. Bellow. Uniform amarts : a class of asymptotic martingales for which almost sure convergence obtains. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 41, 177-191, 1978.

[3] G.A. Edgar. Oral communication, 1997.

[4] G.A. Edgar and L. Sucheston. Stopping times and directed processes. Cambridge University Press, Cambridge, UK, 1992.

[5] L. Egghe. Stopping time techniques for analysts and probabilists. London Mathematical Society Lecture Notes Series 100, Cambridge University Press, UK, 1984.

[6] L. Egghe. Extension of the general "success breeds success" principle to the case that items can have multiple sources. Proceedings of the fifth biennial Conference of the international Society for Scientometrics and Informetrics, Rosary College, River Forest, Ill., USA, 1995. Learned Information, Medford, NJ, 1995.

[7] L. Egghe. Source-item production laws for the case that items have multiple sources with fractional counting of credits. Journal of the American Society for Information Science 47 (10), 730-748, 1996.

[8] L. Egghe and R. Rousseau. Introduction to informetrics. Quantitative methods in library, documentation and information science. Elsevier, Amsterdam, 1990.

[9] L. Egghe and R. Rousseau. Generalized success-breeds-success principle leading to time dependent informetric distributions. Journal of the American Society for Information Science 46, 426-445, 1995.

[10] L. Egghe and R. Rousseau. Stochastic processes determined by a general success-breeds-success principle. Mathematical and Computer Modelling 23(4), 93-104, 1996.

[11] A.G. Mucci. Another martingale convergence theorem. Pacific Journal of Mathematics 62(2), 539-541, 1976.

[12] J. Neveu. Discrete-parameter martingales. North-Holland Publishing Company, Amsterdam, 1975.

[13] R. Rousseau and E. Spinak. Do a field list of internationally visible journals and their journal impact factors depend on the initial set of journals? A research proposal. Journal of Documentation 52(4), 449-456, 1996.