Информационные процессы, Том 15, № 1, 2015, стр. 89–96.

Limit theorems for additive functions of SCOT trajectories

M. Malyutov* **and T. Zhang***

Mathematics Dept., Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA E-mail: m.malioutov@neu.edu, zhang.tong@husky.neu.edu

Received March, 20, 2015

Abstract—Stochastic COntext Tree (abbreviated as SCOT) is m-Markov Chain with every state of a string independent of the symbols in its more remote past than the **context** of **length** determined by the preceding symbols of this state. SCOT has also appeared in other fields under somewhat confusing names VLMC, PST, CTW,... for compression applications. SCOT modeling and its stationary distribution study was the subject of our preceding publication (IP, No.3, 2014). We estimated SCOT parameters and tested homogeneity of data strings using additive functions of SCOT trajectories in IP, No.4, 2013. Here we justify properties of the homogeneity test statistic introduced there and for finding active inputs of sparse systems with correlated noise.

KEYWORDS: variable length Markov chain, stochastic context Tree, asymptotic normality, additive functions of SCOT trajectories, large deviations.

1. INTRODUCTION

Stochastic COntext Tree (abbreviated as SCOT) is m-Markov Chain (m-MC) with every state of a string independent of the symbols in its more remote past than the **context** of **length** determined by the preceding symbols of this state. SCOT has also appeared in other fields under somewhat confusing names VLMC, PST, CTW,... for compression applications. Apparently the first SCOT *Statistical Likelihood* comparison application [1] to *non-stationary Bioinformatics data* seems inadequate. Both asymptotic normality (AN) of additive functions are studied first for ergodic finite alphabet m-MC trajectories. We discuss then what improvements can be obtained for the particular case of sparse SCOT models. Our apparently new results for m-MC follow in straightforward way from those for 1-MC (further called MC). AN for MC is known under appropriate conditions since S.N. Bernstein's works around a century ago and continued in numerous publications. For readers convenience, we refer mostly to the popular comprehensive review [8] available online. We discuss in some detail only Large Deviations (LD) for a particular case of additive functions in discrimination between two composite hypotheses as applied to Separate Testing of Inputs in search of active inputs of a *sparse system with stationary noise*, where error probability under activeness is fixed, while the null hypothesis error should be as small as possible similarly to [5], [4]. More general MC LD results known for almost half a century are surveyed e.g. in [10].

1.1. m-MC reduction to MC on A^m

An m-MC $\{X_n\}$ with a finite state space (alphabet) A can be regarded as 1-MC

$$\{Y_n = (X_n, X_{n+1}, \dots, X_{n+m-1})\}$$

with alphabet as the space of *m-grams* A^m. Namely:

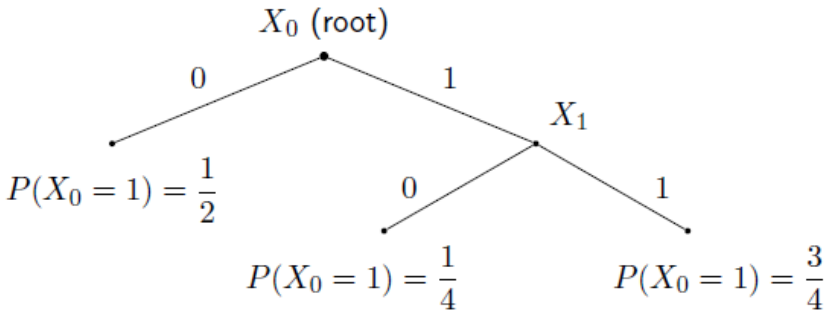

Fig. 1. The simplest stochastic context tree (Model 1).

$P(Y_{n+1}|Y_n) = P(X_{n+m}|Y_n)$, if $X_{n+1}, \ldots, X_{n+m-1}$ coincide in both sides, and 0 otherwise.

This induced MC on A^m is not necessarily ergodic for ergodic m-MC. A simple counterexample follows in section 1.3.

1.2. Simplest SCOT example in Fig.1

- contexts $\{\mathbf{0}\}, \{\mathbf{01}\}, \{\mathbf{11}\}$
- transition probabilities $P(x_0 = 1)$ given preceding contexts are respectively $1/2, 1/4, 3/4$, as displayed above.

1.3. Counterexample

Consider binary 2-MC with alphabet $\{0, 1\}$ and transition probability $1/2$ from $\{0, 1\}$ or $\{1, 1\}$ to $\{0\}$ and $\{1\}$, transition probability 1 from $\{0, 0\}$ and $\{1, 0\}$ to $\{1\}$. It is ergodic, but the induced MC on 2-grams has transient state $\{0, 0\}$.

Although we showed counterexample to availability of SCOT reduction to MC on the space of contexts in [7], this reduction seems to be generally valid as shown by two following examples neither of which satisfies the sufficient condition of [7]. Thus reduction possibility to MC and its ergodicity are either validated or assumed when appropriate.

1.4. 'Comb' Model D_m

Binary Context Tree D_m repeats n times consequently the splitting of the right hand side of the Tree on Fig.1. It has contexts $(0), (01), (011), ..., (01^{m-1}), (1^m)$ and admits reduction to 1-MC for every m. Let us assign all root probabilities of $\{0\}$ and $\{1\}$ given every context as $1/2$.

Then SCOT is ergodic, the **stationary distribution** $\{q_i\}$ for this 1-MC is $(1/2, 1/4, \ldots, (1/2)^{m-1}$, $(1/2)^m, (1/2)^m)$, and entropy rate

$$ER = -\sum_{i \in A} \sum_{j \in A} q_i p_{ij} \log p_{ij}$$

is $\log 2$ for all m.

1.5. Some more SCOT models

Ergodic and non ergodic ladder caricatures of the 'Galileo inertia law with friction' were analyzed in [7]. The motion of Brownian particle hitting heavy objects at random 'Spike' times is described

by another SCOT — Spike model which is defined in two steps. The first step is to assign randomly the increments of the Random walk to *regular ones* with Probability $1 - 2/N$, and (with $2/N$ probability) to spikes. The second step is specifying standard increment distribution for regular increments and SCOT model with increments of magnitude 0, or $\pm\sqrt{N}$ to spikes. Here we sketch the functional convergence proof of the Spike model to a martingale — *mixture of the Brownian motion and a symmetric pair of* $\pm$ *Poisson processes.*

The family $X_n^N = \sum_{i=1}^n r_i$ of Spike models is in few words a reflected Random Walk on large interval $[-l, l], l > N^{3/2}$. Regular part of X_n^N has increments ± 1 and reflects one step from the boundary next time it hits it. Very rare (probability $2/N$) random interruptions by spikes at random Spike time moments n have magnitudes 0 or $\pm\sqrt{N}$ depending on whether $X_n^N = X_{n-1}^N$ (this event can happen only after very unlikely reflection of a spike from the boundary), or $X_n^N > X_{n-1}^N$, or the opposite inequality holds. More formally, $X_n^N = \sum_{i=1}^n r_i, r_i$ *in a regular part* is an equally likely sequence of independent identically distributed (IID) $\pm 1, i = 1, \ldots,$ inside $(-l, l)$, while *irregular part* is a SCOT model specified above.

1.6. Continuous time limit

Let the increments of time/space be respectively $1/N, 1/\sqrt{N}$ instead of 1. Introduce $w^N(t) = N^{-1/2} X^N_{\lfloor Nt \rfloor}$ (summation until the integer part $\lfloor Nt \rfloor$ of Nt). We study the weak convergence of $w^N(t)$ as $N \to \infty$.

Inside $(-l, l)$ *conditionally on no spike at time* $k + 1$

$$E(X_{k+1} - X_k) = 0,$$

$$\mathrm{Var}(X^N_{k+1} - X^N_k)) = (1 - 2/N)/N.$$

Let τ_k be the k-th spike time. Obviously, $\tau_k - \tau_{k-1}$ are IID, independent of σ-algebra spanned by $(x_j, j < k - 2)$ converging to the exponential distribution with mean 2.

Theorem 1. *In the limit we get a weak convergence of* $w^N(t)$ *to the Wiener process* $w(t)$ *in between independent of* $w(t)$ *compound Poisson spikes process of equally likely magnitudes* ± 1*:*

$$P(\tau > t) = \exp(-t/2),$$

τ *and* $\{x_t\}, t < \tau$, *are independent.*

Proof sketch follows in a straightforward way along the lines of the familiar proof of Random Walk weak convergence to the Wiener process, including

- establishing convergence of the Finite Dimensional Distributions (FDD) from the convergence of their multivariate characteristic functions,
- verifying the Kolmogorov Uniform Continuity (KUM) of trajectories criterion *in between spikes* by checking that
$$|w^N(t + h) - w^N(t)|^4 \le const|h|^2,$$
- applying the Prokhorov theorem to the tight family of distributions of X_n^N trajectories.

Denote the events: $\{\pm(k + 1)\} = \{X_k - X_{k-1} = \pm 1\}$. If a spike happens at time $k + 1$, then $X_{k+1} - X_k | \{\pm(k)\} = \pm\sqrt{N}$. As $N \to \infty$, this dependence of the preceding event $X_k - X_{k-1}$ becomes negligible, and the sign of the spike becomes independent of the limiting $w(t)$.

1.7. 'Thorny' $TH_{a,b}$ SCOT model

Our next model is similar to the Spike model, only rare random time moments of spikes $\pm aN^b$ with similar dependence of spikes magnitude on the past take place with probability $N^{-2b}, 0 < b < 1/8$. In the same limiting situation of time intervals $1/N$ and steps $1/\sqrt{N}$, the KUM criterion is valid with similar parameters, thus trajectories of the limiting $T_{a,b}$ model are continuous.

Let the martingale sequence $w^N(t)$ be as above. Then $Er_i = 0, Var[w^N(t)] = Nt[(a^2N^{2b-1})N^{-2b} + (1-N^{-2b})] \to a^2+1$. The equality of summands preceding a spike can be neglected. The covariance of $w^N(t)$ converges to that of $\sqrt{a^2+1}w(t)$ in a similar way. Thus this model gives larger volatility without noticeable drift in the limit to continuous t. The weak FDD convergence to that of $\sqrt{a^2+1}w(t)$ is valid since the Martingale version of the Lindeberg condition holds, see [12]. Thus we proved the following statement.

Proposition. *$w^N(t)$ converges weakly to $\sqrt{a^2+1}w(t)$.*

It looks promising to experiment with fitting discrete time financial data as $TH_{a,b}$-model with variable parameter a estimated as trigonometric series of some order to describe smooth volatility changes.

2. ASYMPTOTIC NORMALITY FOR ADDITIVE FUNCTIONS OF m-MC TRAJECTORIES

Basically, our further limit theorems will be derived for finite m-MC by reducing them to the well-known case of MC on m-grams.

Rates of convergence in these theorems depend on the alphabet size which is significantly lower for sparse SCOT then for general m-MC.

The main steps of our AN straightforward derivation for ergodic m-MC as corollary of that in [8] are:

– Given an ergodic m-MC $\{X_i\}$ with finite alphabet A, denote the induced 1-MC on m-grams (see our Introduction) as $\{Y_i\}$. We assume that MC $\{Y_i\}$ is ergodic which does not generally follows from ergodicity of X_N

– This 1-MC $\{Y_N\}$ is a Harris invariant (see e.g. [8], chapter 17) with respect to a probability distribution.

Let g be a Borel function on $\mathbf{R}$.

– Define $f(Y_i) := f(X_i, X_{i-1}, \dots, X_{i-m+1}) = \sum_{k=0}^{m-1} g(X_{i-k})$.

– Define $\bar{f}_N := (1/N)\sum_{i=1}^{N} f(Y_i), \bar{g}_N := (1/N)\sum_{i=1}^{N} g(X_i)$.

– If π is the stationary distribution and $E_\pi|f^2| < \infty$, then the ergodic theorem ([8], section 17.3) guarantees that $\bar{f}_N \to E_\pi f$ with probability 1 as $N \to \infty$, and the central limit theorem holds for $\bar{f}_N$ ([8], section 17.4):

$\sqrt{N}(\bar{f}_N - E_\pi f) \Rightarrow N(0, f_\pi^2)$ weakly, where $\sigma f_\pi^2 < \infty$ is the variance of f with respect to π.

– $(1/\sqrt{N}\sum_{i=1}^{N} f(Y_i) - E_\pi f) \Rightarrow N(0, \sigma f_\pi^2)$ weakly,

– $(1/\sqrt{N}\sum_{(}\, i=1)^N \sum_{k=0}^{m-1} g(X_{i-k}) - E_\pi f) \Rightarrow N(0, \sigma f_\pi^2)$ weakly,

– $\sqrt{N}(m\bar{g}_N(X) - E_\pi f) \Rightarrow N(0, \sigma f_\pi^2)$ weakly.

– The joint convergence of the sample mean and sample variance to independent respectively Normal and χ^2 distributions is established for ergodic finite MC similarly.

– A sparse ergodic SCOT AN convergence rate under fixed context size is generally better than for the full m-MC. As far as we know, the case of proportional sample and MC alphabet sizes is not yet studied.

– Above results justify t-distribution of our homogeneity test statistic based on studentized averages of SCOT log-likelihoods introduced in section 2.2 .

2.1. Asymptotic expansion for additive functions

In [3] the *first terms* of asymptotic expansion under Cramer-type conditions are:

$$P(N^{-1/2}(\sum_{i=1}^{N} f(x_i) \le x)) = \Phi_\sigma(x) + \phi_\sigma(x)q(x)N^{-1/2} + O(N^{-1}).$$

Here ϕ and Φ are PDF and CDF of the central Normal RV with StD σ, q are expressed in terms of the Hermite polynomial.

Malinovsky finds explicit expression for the polynomial $q(x)$.

This result can be generalized for m-MC by the method displayed above. We believe that the coefficient $q(x)$ for *sparse SCOT* is substantially less than that for general m-MC.

2.2. Nonparametric Homogeneity test

We estimate the SCOT of the large stationary ergodic 'training' string T. Then, using the SCOT of T we first find the loglikelihood $L_Q(k)$ of query slices Q_k and second, of strings S_k simulated from the training distribution of the same size as $Q_k, k = 1, \dots, K$, (for constructing simulated strings, see algorithm in [2]).

We then find log-likelihoods $L_Q(k)$ of Q_k, $L_S(k)$ of S_k using the derived probability model of the training string and the average $\bar{D}$ of their difference D.

Next, due to asymptotic normality of log-likelihood increments, we can compute the usual empirical variance V of $\bar{D}$ and the t-statistic t as the ratio $\bar{D}/\sqrt{V}$ with $K-1$ degrees of freedom (DF). We find K^* from the condition that $t(K^*)$ is maximal. Then, the p-value of homogeneity is evaluated for the t-distribution with $K^* - 1$ DF.

3. EXPONENTIAL TAILS FOR LOG-LIKELIHOOD FUNCTIONS

Introduce diversion (cross entropy) $D(P_1||P_0) = \mathbf{E}_1 \log(P_1/P_0)$ and consider first goodness of fit tests of P_0 vs. P_1 for IID sample of size N.

3.1. 'Stein' lemma for LRT between two known SCOT distributions

(Proved first for IID case by H. Cramer in 1938).

If $D(P_1||P_0) \ge \lambda$ *and any* $0 < \varepsilon < 1$*, then the error probabilities of Likelihood Ratio Test (LRT) satisfy simultaneously*

$$P_0(L_0 - L_1 > N\lambda) \le 2^{-N\lambda}$$

and

$$\lim_{N\to\infty} P_1(L_0 - L_1 > N\lambda)) \geq 1 - \varepsilon > 0.$$

No other test has both error probabilities less in order of magnitude.

3.2. Nonparametric version of the 'Stein' lemma

Generate an artificial N-sequence $\mathbf{z}^N$ independent of $\mathbf{y}^N$ and distributed as P_0 and denote by L_0 its loglikelihood given the SCOT model of the training string.

L is the query log-likelihood given the SCOT model of the training string.

Also assume that the joint distribution of S slices of size N converge to their product distribution in Probability.

Theorem 2. *Suppose P_1, P_0 are SCOT, $D(P_1||P_0) > \lambda$ and we reject homogeneity, if the 'conditional version of the Likelihood Ratio' test $\mathcal{T} = \overline{L} - \overline{L}_0 > \lambda$. Then the same error probability asymptotic as for LRT in the 'Stein' lemma is valid for this test.*

Proof sketch. Under negligible brakes and independent slices, their probabilities multiply. To transparently outline our ideas (with some abuse of notation) replace the condition under summation sign to a similar one for the whole query string: instead of $P_0(T' > 0) = \sum_{\mathbf{y},\mathbf{z}:\overline{L}-\overline{L}_0>\lambda} P_0(\mathbf{y})P_0(\mathbf{z})$, we write the condition under the summation as $L - L_0 > N\lambda$ which is approximated in Probability under P_0 by $L_N(\mathbf{y}) - |\mathbf{z}| > N\lambda$.

Thus

$$P_0(T' > 0) \leq \sum_{\mathbf{z}} \sum_{\mathbf{y}:\ P_0(\mathbf{y})\leq 2^{-N\lambda-|\mathbf{z}|}} P_0(\mathbf{y})P_0(\mathbf{z}) \leq 2^{-N\lambda} \sum_{\mathbf{z}} 2^{-|\mathbf{z}|} = 2^{-N\lambda}$$

by the Kraft inequality. We refer to [11] for an accurate completion of our proof in a similar situation.

Informally again,

$$\lim_{N\to\infty} P_1(T' > 0) = \lim_{N\to\infty} P_1(N^{-1}(|\mathbf{y}| - |\mathbf{z}|) > \lambda) = D(P_1||P_0) + \varepsilon, \varepsilon > 0.$$

$|\mathbf{y}|/N$ is in Probability P_1 around $-\log P_0(\mathbf{y}) = E_1(-\log(P_0(\mathbf{y}))) + r$, $|\mathbf{z}|/N$ is in Probability P_0 around $-\log P_0(\mathbf{z}) = h_0^N + r'$. As in the Consistency proof, all the principal deterministic terms drop out, and we are left with the condition $r < \varepsilon + r'$ which probability converges to 1 since both r, r' shrink to zero in the product $(\mathbf{y}, \mathbf{z})$-Probability as $N \to \infty$.

4. APPLICATION TO THE STI ANALYSIS UNDER COLORED NOISE

Suppose that we can assign an arbitrary t-tuple of binary inputs $\mathbf{x} := (x(j), j \in [t]), [t] := \{1, \ldots, t\}$, and measure the **noisy output** Z in a measurable space $\mathcal{Z}$ such that $P(z|y)$ is its conditional distribution given *'intermediate output'* $y = g(x(A))$ with a finite-valued function $g(\cdot)$, and $P(z|x(A))$ is their superposition (Multi Access Channel (MAC)), where $x(A)$ is an s-tuple of of 'Active Inputs' (as in [5]). Conversely, every MAC can be decomposed uniquely into such a superposition. We omit obvious generalization to q-ary inputs.

The Mean Error probability (MEP) $\gamma > 0$ is the probability of a test T based on a sequence of measurements $\mathbf{z} \in \mathcal{Z}^N$ to misidentify s-tuple A over the direct product of uniform prior for A and noise distribution of $\mathbf{z}$.

It is well-known that for **known MAC** $P(z|x(A))$, the Maximum Likelihood (ML)-decision minimizes the MEP for any design. It generalizes the Brute Force analysis of noiseless data for the case of known MAC distribution. If MAC is unknown, a **universal nonparametric test** of computational complexity $O(t\log t)$ for IID noise and a random design is $\mathcal{I}$ ([5]).

The universal STI decision chooses maximal values of the $\mathcal{I}$ for a-th input and output, $a=1,\dots,t$. $\mathcal{I}$ for IID noise is the Empirical Shannon Information (ESI):

$$\mathcal{I}(\tau_N^N(A)) = \sum_{x(a)\in\mathcal{B}^{|A|}} \sum_{z\in\mathcal{Z}^{\mathcal{N}}} \tau(x(a),z)\log(\tau(z|x(a))/\tau(z)).$$

Let us now consider 'Colored' noise. Given arbitrary intermediate vector–output $\mathbf{y}$, the sequence $\mathbf{z}$'s conditional distribution is that of a *stationary ergodic SCOT random string* taking values from a finite alphabet $\mathcal{Z}^N$.

The STI-test for colored noise is defined as follows:

Denote

$$\mathcal{U} = \mathcal{B}\times\mathcal{Z},$$

and consider for a given $j=1,\dots,N$ two N-sequences with letters from $\mathcal{U}$:

$$u_j{}^N := (x_j(i),z(i)), i=1\dots,N,$$

and

$$v_j^N := (x_j(i)(\times)z(i)), i=1\dots,N,$$

taken from the original and generated product-distributions, digitize them into binary sequences $\mathbf{U}_j^M, \mathbf{V}_j^M$ of appropriate length and evaluate the SCOT homogeneity statistic (see further) of the product P_0 and original distributions $P_1 = P_1^j$. The SCOT trained loglikelihoods of these bi-variate strings are $\log P(\mathbf{V}^M) = L^M$ — the main inference tool about P_0.

Consider a **query** binary sequence $\mathbf{U}^M$ distributed as P_1 and test whether the homogeneity hypothesis $P_0 = P_1$ contradicts the data or not. Let us partition $\mathbf{y}^M$ into several **slices** $\mathbf{U}_i, i=1,\dots,S$, of identical length N divided by 'brakes' of length $2m$ which are sufficient to ensure independence of slices for m-MC. Introduce strings $\mathbf{C}_i = (\mathbf{V}^M, \mathbf{U}_i)$. Define L_i — statistic and $\overline{L}$ = average of all L_i. Similarly, $\overline{L}_0$ = average of all L_{0i} with $\mathbf{U}_i$ replaced with independent P_0- distributed slices of the same length. Finally, homogeneity statistic is $\bar{R} = \overline{L} - \overline{L}_0$. The $\bar{R}$ test is shown in our Theorem 2 to have the same exponential tail under P_0 as the asymptotically optimal Likelihood Ratio test, if the error probability under alternative is arbitrarily small but positive and fixed which is natural, if $s=$ const, $t\to\infty$. This fact can be used to prove that the STI test is asymptotically optimal among all tests of affordable computational complexity $O(t\log t)$ based on separate influence comparison of inputs on the output of the system with stationary noise under a random design.

5. CONCLUSIONS, DISCUSSION AND FUTURE DIRECTIONS

We derive asymptotic normality of additive functions of SCOT trajectories to apply for verifying the t-distribution of the homogeneity test statistic applied in [7] for many real data case studies. We prove the optimality of exponential tails of our homogeneity test and apply it for showing asymptotic optimality of our nonparametric STI test applicable for searching active inputs of a system with colored noise. We introduce several new SCOT models and study the functional

limit theorems under 'continuous time' conditions. These results suggest experimenting with semi-parametric trigonometric regression of the 'Thorny' model for modeling financial time series with regularly varying volatility for comparison with the well-known GARCH approach.

Acknowledgement P. Grosu and Prof. I. Tsitovich generously aided us in the final technical preparation of this paper (and of our previous papers published in IP), V. Rotar's clarification helped proving our Proposition.

REFERENCES

1. Bejerano G. *Automata learning and stochastic modeling for biosequence analysis.* PhD dissertation, Hebrew University, Jerusalem, 2003.
2. Mächler M. and Bühlmann P. Variable Length Markov Chains: methodology, computing, and software, *Journal of Computational and Graphical Statistics*, 2004, Vol. 13, No. 2, 435–455.
3. Malinovsky V.K. On limit theorems for Harris Markov chains. I. *Theory Probab. Appl., (English)* 1987, Vol 31, 269–285.
4. Malyutov M.B. Compression based homogeneity testing. *Doklady of Russian Acad. Sci.*, 2012, Vol. 443, No. 4, 427–430.
5. Malyutov M.B. Search for active inputs of a sparse system: a review, *Springer Lecture Notes in Computer Science*, Vol. 7777, 478–507, 2012.
6. Malyutov M., Zhang T., Li Y., and Li X. Time series homogeneity tests via VLMC training. *Information Processes*, 2013, Vol. 13, No. 4, 401–414.
7. Malyutov M., Grosu P., and Zhang T. SCOT stationary distribution evaluation for some examples. *Information Processes*, 2014, Vol. 14, No. 3, 275–283.
8. Meyn S.P. and R.L. Tweedy R.L. *Markov chains and stochastic stability.* Springer, 1993.
9. Rissanen J. A universal data compression system. *IEEE Trans. Inform. Theory*, 1983, Vol. 29, No. 5, 656–664.
10. Rassoul-Agha F. and Seppalainen T. *A Course on Large Deviations with an Introduction to Gibbs Measures*, Department of Mathematics, University of Utah, Salt Lake City, 2010.
11. Ziv J. On classification and universal data compression. *IEEE Trans. on Inform. Theory*, 1988, Vol. 34, No. 2, 278–286.
12. Liptser R. Sh., Shiryayev A.N. *Theory of martingales.* Mathematics and its Applications (Soviet Series), Kluwer Academic Publishers Group, Dordrecht, Vol 49, 1989.