# DEVIATION PROBABILITY BOUND FOR MARTINGALES WITH APPLICATIONS TO STATISTICAL ESTIMATION

LIPTSER, R. AND SPOKOINY, V.

*Dept. Electrical Engineering-Systems,*
*Tel Aviv University,*
*69978 Tel Aviv, Israel*

*and*

*Weierstrass Institute for Applied Analysis and Stochastics,*
*Mohrenstr. 39, 10117 Berlin, Germany*

1991 *Mathematics Subject Classification.* 62G05; Secondary 62M99.

*Key words and phrases.* martingale, deviation probability, maximum likelihood estimate, autoregression, linear diffusion.

ABSTRACT. Let $M_t$ be a vector martingale and $\langle M\rangle_t$ denote its predictable quadratic variation. In this paper we present a bound for the probability that $z^*\langle M\rangle_t^{-1}M_t > \lambda\sqrt{z^*\langle M\rangle_t^{-1}z}$ with a fixed vector $z$ and discuss some its applications to statistical estimation in autoregressive and linear diffusion models. Our approach is non-asymptotic and does not require any ergodic assumption on the underlying model.

## 1. INTRODUCTION. STATISTICAL EXAMPLES

Let observations $Y_1, \dots, Y_T$ be generated by the linear regression model:

$$Y_t = X_t^*\theta + \varepsilon_t, \qquad t = 1, \dots, T, \tag{1.1}$$

where $\theta \in \mathbb{R}^p$ is unknown vector of parameters, $X_t$, $t = 1, \dots, T$, are deterministic design points from $\mathbb{R}^p$, and $(\varepsilon_t)_{t\geq 1}$ is a sequence of i.i.d. zero mean Gaussian random variables with the variance $\sigma^2$. Hereafter, all vectors are assumed to be vector-columns and $a^*$ (resp. $\|a\|$) means the transpose (resp. the Euclidean norm) of the vector $a$.

For estimating the vector $\theta$, one usually applies the maximum likelihood estimate (MLE) $\widehat{\theta}$:

$$\widehat{\theta} = \left(\sum_{t=1}^T X_t X_t^*\right)^{-1} \sum_{t=1}^T X_t Y_t. \tag{1.2}$$

(the matrix $\sum_{t=1}^T X_t X_t^*$ is assumed to be non singular). The estimation error

$$\widehat{\theta} - \theta = \left(\sum_{t=1}^T X_t X_t^*\right)^{-1} \sum_{t=1}^T X_t(Y_t - X_t^*\theta) = \left(\sum_{t=1}^T X_t X_t^*\right)^{-1} \sum_{t=1}^T X_t\varepsilon_t \tag{1.3}$$

is a zero mean Gaussian vector. Its covariance matrix, which is often called often the *information matrix*, reads as follows:

$$W = \boldsymbol{E}(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^* = \sigma^2 \left(\sum_{t=1}^T X_t X_t^*\right)^{-1}$$

By $w_{k,k'}$, $k, k' = 1, \dots, p$ we denote the elements of the matrix $W$. The property $\widehat{\theta} - \theta \sim \mathcal{N}(0, W)$ implies: for every $\lambda > 0$ and $k = 1, \dots, p$

$$\boldsymbol{P}\left(|\widehat{\theta}_k - \theta_k| > \lambda w_{kk}^{1/2}\right) \leq 2e^{-\frac{\lambda^2}{2}}. \tag{1.4}$$

The aim of this paper is to establish a similar exponential bound for probability of deviations $\widehat{\theta} - \theta$ for more complicated statistical models arising in time series analysis. Below we present two typical examples.

*Example* 1.1. [Autoregression model] Let observations $Y_1, Y_2, \dots, Y_T$ follow the autoregression equation

$$Y_t = \theta_1 Y_{t-1} + \dots + \theta_p Y_{t-p} + \varepsilon_t, \tag{1.5}$$

where one sets $Y_0, Y_{-1}, \dots, Y_{1-p} = 0$ and $(\varepsilon_t)_{t\geq 1}$ are i.i.d. Gaussian random errors with parameters $(0, \sigma^2)$.

Introduce a vector $\theta$ of the unknown coefficients $\theta = (\theta_1, \dots, \theta_p)^* \in \mathbb{R}^p$ and define $X_t = (Y_{t-1}, \dots, Y_{t-p})^* \in \mathbb{R}^p$. Then, the original autoregression equation given in (1.5) admits the 'regression-like' representation (compare (1.1)):

$$Y_t = X_t^* \theta + \varepsilon_t.$$

Moreover, formula (1.2) (resp. (1.3)) for the MLE $\widehat{\theta}$ (resp. for the deviation $\widehat{\theta} - \theta$) remains valid for the autoregression case as well. Despite of this similarity, there is an essential difference between regression and autoregression models. For the autoregression case, the 'design' points $X_1, X_2, \dots$ are random and heavy correlated with the observations $Y_1, Y_2, \dots$. Therefore, the matrix $W = \left(\sum_{t=1}^T X_t X_t^*\right)^{-1}$, which is often called the *conditional covariance* or *conditional information matrix*, is also random and heavy correlated with the observations. Hence, the estimation error $\widehat{\theta} - \theta$ is no more a Gaussian vector and the bound (1.4) does not apply.

To analyze properties of the deviation $\widehat{\theta} - \theta$ for this situation, introduce a valued in $\mathbb{R}^p$ process

$$M_t = \sum_{s=1}^t X_s \varepsilon_s, \qquad t \geq 1.$$

Since $X_t$ depends only on $Y_1, \dots, Y_{t-1}$, and since $\varepsilon_t$ is independent of $Y_1, \dots, Y_{t-1}$, the process $(M_t)_{t\geq 1}$ is a vector square integrable martingale with respect to the filtration generated by $(\varepsilon_t)_{t\geq 1}$. The predictable quadratic variation of this martingale reads as follows

$$\langle M \rangle_t = \sigma^2 \sum_{s=1}^t X_s X_s^*, \ t \geq 1,$$

so that $W = \langle M \rangle_T$. With this notation, on the set where $\langle M \rangle_T$ is non singular, we have

$$\widehat{\theta} - \theta = \langle M \rangle_T^{-1} M_T.$$

Therefore, the original statistical problem leads to evaluation of

$$\boldsymbol{P}\left(|z^* \langle M \rangle_T^{-1} M_T| > \lambda \sqrt{z^* \langle M \rangle_T^{-1} z}\right) \tag{1.6}$$

where $z$ is a deterministic vector.

*Example* 1.2 (Diffusion model). Let the observed process $X_t$ follow the Itô equation (with respect to Wiener process $w_t$)

$$\mathrm{d}X_t = \theta^* f_t \, \mathrm{d}t + \sigma_t \, \mathrm{d}w_t, \qquad X_0 = 0. \tag{1.7}$$

Here $\theta \in \mathbb{R}^p$ is an unknown vector, $f_t \in \mathbb{R}^p$ and $\sigma_t \in \mathbb{R}_+$ are observed random processes such that for every $t > 0$, it holds $\int_0^t \|f_s\|^2 \sigma_s^{-2}\,\mathrm{d}s < \infty$. The particular cases of (1.7) are: the Orstein-Uhlenbeck model ($p = 1$)

$$\mathrm{d}X_t = \theta X_t\,\mathrm{d}t + \mathrm{d}w_t,$$

a nonlinear autoregression model

$$\mathrm{d}X_t = \theta h(X_t)\,\mathrm{d}t + s(X_t)\,\mathrm{d}w_t$$

and a model with delay, when $h(X_t)$ and $s(X_t)$ are replaced by $h(X_{t-\Delta})$ and $s(X_{t-\Delta})$, $\Delta$ being the delay parameter.

The MLE estimate $\widehat{\theta}$ of $\theta$ from (1.7) reads as follows:

$$\widehat{\theta} = \left(\int_0^T f_t f_t^* \sigma_t^{-2}\,\mathrm{d}t\right)^{-1} \int_0^T f_t \sigma_t^{-2}\,\mathrm{d}X_t$$

so that the error of estimation $\widehat{\theta} - \theta$ can be represented in the form

$$\widehat{\theta} - \theta = \left(\int_0^T f_t f_t^* \sigma_t^{-2}\,\mathrm{d}t\right)^{-1} \int_0^T f_t \sigma_t^{-2}(\,\mathrm{d}X_t - f_t^*\theta\,\mathrm{d}t) = \langle M\rangle_T^{-1} M_T, \tag{1.8}$$

where

$$M_t = \int_0^t f_s \sigma_s^{-1}\,\mathrm{d}w_s, \qquad t \geq 0,$$

is a continuous vector martingale and

$$\langle M\rangle_t = \int_0^t f_s f_s^* \sigma_s^{-2}\,\mathrm{d}s \tag{1.9}$$

is its predictable quadratic variation.

We see that for both examples, the study of the properties of the MLE $\widehat{\theta}$ leads to establishing a proper bound for probability of the form (1.6).

Some other examples where similar problems arise can be found in Liptser and Spokoiny (1997) in context of adaptive nonparametric estimation of the drift function for two-scaled diffusion systems and in Härdle, Spokoiny and Teyssière (1999) for estimation of parameters for time inhomogeneous financial data.

The majority of general martingale results (see e.g. Liptser and Shiryaev (1986), Jacod and Shiryaev (1987)) concern only with asymptotic properties of $M_T$, as $T \to \infty$, under some conditions on the behaviour of $\langle M\rangle_T$. Particularly, if for some deterministic factors $b_T \to 0$ as $T \to \infty$, random matrices $b_T\langle M\rangle_T$ converge to a non singular deterministic matrix $\Sigma$, and also, for the discrete time case, the Lindeberg condition holds: for every $\varepsilon > 0$

$$\lim_{T\to\infty} b_T \boldsymbol{E} \sum_{t=1}^{T} \bigl(M_t - M_{t-1}\bigr)^2 I(|M_t - M_{t-1}| > \varepsilon) = 0,$$

then $b_T^{1/2} M_T$ is asymptotically, as $T \to \infty$, normal with zero mean and the covariance matrix $\Sigma$ and the bound (1.4) holds in the following asymptotic sense ($\widehat{\theta_k} = \widehat{\theta}_k(T)$, $w_{kk}^{1/2} = w_{kk}^{1/2}(T)$): for every fixed $\lambda > 0$

$$\overline{\lim_{T\to\infty}} \boldsymbol{P}\left(|\widehat{\theta}_k(T) - \theta_k| > \lambda w_{kk}^{1/2}(T)\right) \le 2e^{-\frac{\lambda^2}{2}}. \tag{1.10}$$

If $b_T\langle M\rangle_T$ converges in probability to a random matrix $\Sigma$, then the vector $b_T^{1/2} M_T$ is asymptotically mixed normal in the sense that the pairs $(b_T^{1/2} M_T, b_T\langle M\rangle_T)$ converge in distribution to the pair $(\Sigma^{1/2}U, \Sigma)$ where $U$ is an independent of $\Sigma$ standard Gaussian vector (see, e.g. Liptser and Shiryaev, 1988, Ch. 5). This again leads to the same asymptotic statement as in (1.10). Unfortunately, these results hold only under rather strong conditions on asymptotic behaviour of $\langle M\rangle_T$ as $T \to \infty$ and do not serve effectively the case of a finite $T$ or a large $\lambda$.

In the case of a scalar unknown parameter, the time-scale arguments, see e.g. Rootzen (1983), help to get some non-asymptotic results but only for the case of scalar parameter $\theta$ and for specially introduced random time moments $T$. An application of this idea to statistical problems for autoregressive and diffusion models leads to the so called sequential estimation, when the underlying parameter is estimated from the sample $Y_1, \dots, Y_\tau$ with a specially defined stopping time $\tau$, see e.g. Novikov (1972) for the case of a linear diffusion model and Grambsch (1983), Lai and Siegmund (1983), Shiryaev and Spokoiny (1997) for the Ornstein-Uhlenbeck model. Some generalizations to the vector autoregression in the special context of guaranteed estimation can be found in Konev and Pergamanshchikov (1996).

There exists also vast literature devoted specifically to the problem of estimating the parameter $\theta$ for autoregressive and linear diffusion models. Here again, the asymptotic approach based on a preliminary study of asymptotic properties of the process $\langle M\rangle_t$ as $t \to \infty$, is usually used. For instance, for the first order autoregression (1.6), one distinguishes between three essentially different cases depending on the value of the unknown parameter $\theta_1$: *ergodic* for $|\theta_1| < 1$, *unstable* for $|\theta_1| = 1$ and *explosive* for $|\theta_1| > 1$. In the ergodic case, the quantity $T^{-1}\langle M\rangle_T = T^{-1}\sum_{t=1}^T Y_{t-1}^2$ converges to a fixed value and the MLE is asymptotically normal. For $|\theta_1| > 1$, the quadratic variation $\langle M\rangle_T$ grows exponentially with $T$ so that $e^{-2T|\theta_1|}\langle M\rangle_T$ converges in probability to some random variable $\Sigma$. The sums $M_T = \sum_{t=1}^T Y_{t-1}\varepsilon_t$ normalized by $e^{T|\theta_1|}$, turns out to be asymptotically mixed normal in the sense

$$\left(e^{-T|\theta_1|} M_T, e^{-2T|\theta_1|}\langle M\rangle_T\right) \xrightarrow{\text{w}} (\Sigma^{1/2}U, \Sigma)$$

where $U$ is standard normal and independent of $\Sigma$. Hence, the normalized estimation error $e^{T|\theta_1|}(\widehat{\theta}(T) - \theta) = e^{T|\theta_1|}\langle M\rangle_T^{-1} M_T$ is also asymptotically mixed normal and the bound (1.4) applies in the asymptotic case, see White (1958). But for $|\theta_1| = 1$, the quadratic variation $\langle M\rangle_T$ grows as $T^2$ in the sense that $T^{-2}\langle M\rangle_T$ converges in law

to some non degenerated distribution, and the deviation $T(\widehat{\theta} - \theta)$ weakly converges to some special law which is neither normal nor mixed normal. Similar results for the autoregression of order $p > 1$ can be found in Basawa and Scott (1983), Chan and Wei (1988), Jeganathan (1988) or Cox and Llatas (1991).

In this paper, we aim to state an exponential upper bound for the probability from (1.6) for a general vector case and in the non asymptotic set-up. This, of course, makes the problem much more complicated and in particular, we are not able to establish the required bound exactly in the form given in (1.4). Our basic result, presented in the next section, describes a bound of the following type

$$\boldsymbol{P}\left( |z^* \langle M \rangle_T^{-1} M_T| > \lambda \sqrt{z^* \langle M \rangle_T^{-1} z},\ \langle M \rangle_T^{-1} \text{ is non singular} \right) \leq P(\lambda) e^{-\lambda^2/2}$$

where $P(\lambda)$ is a polynomial of the degree $p$ whose coefficients are connected to regularity conditions on the matrix $\langle M \rangle_T$.

Section 3 contains some statistical applications.

## 2. Deviation probability for martingales

Let $U$ be a zero mean Gaussian random vector valued in $\mathbb{R}^p$ with a positively definite covariance matrix $V$: $\boldsymbol{E}U = 0$, $\boldsymbol{E}UU^* = V$. Then $V^{-1}U$ is also a Gaussian random vector with parameters $(0, V^{-1})$. In particular, for every fixed vector $z \in \mathbb{R}^p$, the scalar product $z^* V^{-1} U$ is a zero mean Gaussian random variable with the variance $z^* V^{-1} z$ and therefore

$$\boldsymbol{P}\left( |z^* V^{-1} U| > \lambda \sqrt{z^* V^{-1} z} \right) \leq 2 e^{-\frac{\lambda^2}{2}}, \qquad \lambda > 0.$$

In this section, we present a similar result for a random non Gaussian vector $U$. More precisely, given a square integrable vector martingale $(M_t)_{t \geq 0}$ with $M_0 = 0$ ($\langle M \rangle_t$, $t \geq 0$, denotes its predictable quadratic variation), we establish an exponential upper bound for the probability of the event

$$\{ z^* \langle M \rangle_T^{-1} M_T > \lambda \sqrt{z^* \langle M \rangle_T^{-1} z},\ \langle M \rangle_T \text{ is nonsingular} \}.$$

We consider here two different cases. The first one corresponds to discrete time martingales with conditionally Gaussian increments while the second one concerns with continuous martingales.

### 2.1. The model in discrete time.

Let $M = (M_t)_{t \in \mathbb{N}}$, $\mathbb{N} = \{0, 1, 2, \dots\}$, be a square integrable martingale with $M_0 = 0$, valued in $\mathbb{R}^p$, $p \geq 1$, defined on a probability space $(\Omega, \mathcal{F}, \boldsymbol{P})$ supplied with filtration $\mathbb{F} = (\mathcal{F}_t)_{t \in \mathbb{N}}$ (i.e. $\boldsymbol{E}(M_t \mid \mathcal{F}_{t-1}) = M_{t-1}$ and $\boldsymbol{E}\|M_t\|^2 < \infty$ for all $t \in \mathbb{N}$). The predictable quadratic variation $\langle M \rangle$ of $M$ is defined via increments $\xi_t = M_t - M_{t-1}$:

$$\begin{aligned}\Sigma_t &= \boldsymbol{E}\left(\xi_t\xi_t^* \mid \mathcal{F}_{t-1}\right),\\ \langle M\rangle_t &= \sum_{s=1}^{t}\Sigma_s.\end{aligned}$$

Obviously, $\langle M\rangle_t$ is the predictable random process (i.e. $\langle M\rangle_t$ is $\mathcal{F}_{t-1}$ measurable) valued in the set of $p\times p$ symmetric non negatively definite matrices (for more details see e.g. Liptser and Shiryaev [13], Ch.1 §8). Our main assumption is that for each $t$, the increment $\xi_t = M_t - M_{t-1}$ is conditionally, given $\mathcal{F}_{t-1}$, Gaussian random vector with conditional parameters $(0,\Sigma_t)$: for every $\gamma\in\mathbb{R}^p$ and $t\geq 1$

$$\boldsymbol{E}\left(e^{\gamma^*\xi_t}\Big|\mathcal{F}_{t-1}\right) = \exp\left(\frac{1}{2}\gamma^*\Sigma_t\gamma\right)\qquad \boldsymbol{P}-\text{a.s.} \tag{2.1}$$

Note that (2.1) does not imply that $M$ is a Gaussian process. A specific example of a martingale, obeying (2.1), is delivered by autoregressive processes from Example 1.1. The condition (2.1) implies that the process

$$Z_t(\gamma) = \exp\left(\gamma^*M_t - \frac{1}{2}\gamma^*\langle M\rangle_t\gamma\right),\ t\in\mathbb{N}$$

is a martingale. In fact,

$$Z_t(\gamma) = Z_{t-1}(\gamma)\exp\left(\gamma^*\xi_t - \frac{1}{2}\gamma^*\Sigma_t\gamma\right)$$

and (2.1) provides $\boldsymbol{E}(Z_t(\gamma)|\mathcal{F}_{t-1}) = Z_{t-1}(\gamma)$, $\boldsymbol{P}$-a.s. Hence $\boldsymbol{E}Z_t(\gamma) = 1$ for every $t\in\mathbb{N}$. This also implies for every stopping time $T$

$$\boldsymbol{E}Z_T(\gamma)\leq 1 \tag{2.2}$$

see Problem 1.4.4. in Liptser and Shiryaev [13].

2.2. **The model in continuous time.** Let $M = (M_t)_{t\in\mathbb{R}_+}$ be a continuous vector martingale in $\mathbb{R}^p$ with $M_0 = 0$, defined on a probability space $(\Omega,\mathcal{F},\boldsymbol{P})$ supplied with filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ complying with, so called general conditions, see Liptser and Shiryaev [13], Ch.1. By $\langle M\rangle = (\langle M\rangle_t)_{t\geq 0}$ we denote the predictable quadratic variation of $M$, see again [13], Ch.1 §1 and §8). As in the discrete time case, introduce the positive process

$$\mathfrak{Z}_t(\gamma) = \exp\left(\gamma^*M_t - \frac{1}{2}\gamma^*\langle M\rangle_t\gamma\right).$$

By the Itô formula $d\mathfrak{Z}_t(\gamma) = \mathfrak{Z}_t(\gamma)\gamma^*dM_t$, and hence the process $\mathfrak{Z}_t$ is a continuous positive local martingale and simultaneously, by Problem 1.4.4. in Liptser and Shiryaev [13]), a supermartingale. Due to the supermartingale property, for every stopping time $T$

$$\boldsymbol{E}\mathfrak{Z}_T(\gamma) \le 1. \tag{2.3}$$

2.3. **Bound for scalar martingale.** We first examine the case when $(M_t)_{t\ge 0}$ is a scalar martingale. Since the proof is based only on (2.2) and (2.3), we do not specify here whether $t$ runs over $\mathbb{N}$ or $\mathbb{R}_+$.

The result is of independent interest and it will be essentially used when studying the general vector case.

**Theorem 2.1.** *Let $T$ be fixed or stopping time. For every $b>0$, $S\ge 1$ and $\lambda \ge 1$*

$$\boldsymbol{P}\left(|M_T| > \lambda\sqrt{\langle M\rangle_T},\ b\le \sqrt{\langle M\rangle_T}\le bS\right) \le 4\sqrt{e}\lambda\,(1+\log S)\,e^{-\frac{\lambda^2}{2}}.$$

*Proof.* The statement follows from

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ b\le \sqrt{\langle M\rangle_T}\le bS\right) \le 2\sqrt{e}\lambda\,(1+\log S)\,e^{-\frac{\lambda^2}{2}} \tag{2.4}$$

and from the similar result for $-M_T$. So, it suffices to check (2.4) only.

Given $a>1$, introduce the geometric series $b_k = ba^k$ and define random events $\mathcal{C}_k = \{b_k \le \sqrt{\langle M\rangle_T} < b_{k+1}\}$, $k=0,1,\dots,K$, where $K$ stands for the integer part of $\log_a S$. Obviously

$$\begin{aligned}&\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ b\le \sqrt{\langle M\rangle_T}\le bS\right) \\ &\qquad \le \sum_{k\ge 0}^{K} \boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ b\le \sqrt{\langle M\rangle_T}\le bS,\ \mathcal{C}_k\right).\end{aligned} \tag{2.5}$$

For every $\gamma$, (2.2) (or (2.3)) implies

$$\boldsymbol{E}\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\ \mathcal{C}_k\right)\exp\left(\gamma M_T - \frac{\gamma^2}{2}\langle M\rangle_T\right) \le 1.$$

Next, taking $\gamma_k = \frac{\lambda}{b_k}$, we obtain

$$\begin{aligned}1 \ &\ge\ \boldsymbol{E}\exp\left(\frac{\lambda}{b_k}M_T - \frac{\lambda^2}{2b_k}\langle M\rangle_T\right)\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right) \\ &\ge\ \boldsymbol{E}\exp\left(\frac{\lambda^2}{b_k}\sqrt{\langle M\rangle_T} - \frac{\lambda^2}{2b_k}\langle M\rangle_T\right)\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right) \\ &\ge\ \boldsymbol{E}\exp\left\{\inf_{b_k\le v\le b_{k+1}}\left(\frac{\lambda^2 v}{b_k} - \frac{\lambda^2 v^2}{2b_k^2}\right)\right\}\boldsymbol{I}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right)\end{aligned}$$

and, since "$\inf_{b_k\le v\le b_{k+1}}$" is attained at the point $v = b_{k+1} = ab_k$, we end up with

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T}, \mathcal{C}_k\right) \le \exp\left\{-\lambda^2\left(a - \frac{a^2}{2}\right)\right\}.$$

Inserting this bound in (2.5) and using that $K \le \log_a S$, we get

$$\boldsymbol{P}\left(M_T > \lambda\sqrt{\langle M\rangle_T},\, b \le \sqrt{\langle M\rangle_T} \le bS\right) \le (1+\log_a S)\exp\left\{-\lambda^2\left(a-\frac{a^2}{2}\right)\right\}.$$

Finally, since the left side of this inequality does not depend on $a$, we may pick $a$ to make the right side possibly small. This leads to the choice $a = 1 + 1/\lambda$ so that

$$\lambda^2\left(a-\frac{a^2}{2}\right) = \lambda^2\left\{1+\frac{1}{\lambda}-\frac{1}{2}\left(1+\frac{1}{\lambda}\right)^2\right\} = \frac{1}{2}(\lambda^2-1).$$

Since also $\log(1+1/\lambda) \ge 1/(2\lambda)$ for $\lambda \ge 1$, we obtain $\log_a S \le 2\lambda \log S$ and (2.4) follows. □

2.4. **Bound for vector martingale.** For the convenience of notation, set $p = d+1$ so that we consider martingale $M = (M_t)$ valued in $\mathbb{R}^{d+1}$, $d \ge 1$. Let $T$ be fixed or stopping time. Define $V = \langle M\rangle_T$ and let $W$ stand for the inverse matrix of $V$ on the set, where $V$ is non singular, $W = \langle M\rangle_T^{-1}$. We deal with the random vector

$$U = WM_T \quad \big(= \langle M\rangle_T^{-1} M_T\big).$$

Hereafter, the elements of the matrix $W$ (resp. of the vector $U$) are denoted by $w_{ij}$, $i,j = 0,\dots,d$ (resp. $U_i$, $i = 0,\dots,d$). Given a vector $z$ from $\mathbb{R}^{d+1}$, we establish an upper bound for the probability of the event $\{|z^*U| > \lambda\sqrt{z^*Wz}\}$ restricted to a set $\mathfrak{A}$, where the matrix $V$ satisfies some regularity conditions given below. We start with the vector $z$ of the form $z = (1,0,\dots,0)^*$ and postpone the general case until Subsection 2.5.

With the specified $z$ we have

$$\{|z^*U| > \lambda\sqrt{z^*Wz}\} = \{|U_0| > \lambda\sqrt{w_{00}}\}.$$

For some positive constants $b$, $S$ $\rho$, $r$, define

$$\mathfrak{A} = \left\{ \begin{array}{c} b \le w_{00}^{-1} \le bS, \\ \\ w_{00}\|V\|_\infty \le r, \\ \\ |w_{0k}/w_{00}| \le \rho, \quad \forall k = 1,\dots,d \end{array} \right\},$$

where $\|V\|_\infty = \sup_{\{\mu\in\mathbb{R}^{d+1}:\|\mu\|=1\}} \|V\mu\|$ is the norm of the matrix $V$.

In many cases, the values $b$, $S$, $\rho$ and $r$ can be chosen such that the probability of $\mathfrak{A}$ is closed to 1 for sufficiently large $T$, see Subsection 2.6.

**Theorem 2.2.** *Let $T$ be fixed or stopping time. For every $b > 0$, $S \geq 1$ $\rho > 0$, $r \geq 1$, and $\lambda \geq \sqrt{2}$*

$$\boldsymbol{P}\left(|U_0| > \lambda\sqrt{w_{00}}, \mathfrak{A}\right) \leq 4e\log(4S)\left(1 + 2\rho\sqrt{rd}\,\lambda\right)^d \lambda e^{-\frac{\lambda^2}{2}}.$$

*Proof.* Set $v_k = w_{0,k}/w_{00}$, $k = 1, \ldots, d$. On the set $\mathfrak{A}$, we have $|v_k| \leq \rho$. Define the random vector $v = (1, v_1, \ldots, v_d)^*$ and note that

$$\boldsymbol{P}\left(|U_0| > \lambda\sqrt{w_{00}}, \mathfrak{A}\right) = \boldsymbol{P}\left(|v^* M_T| > \lambda\sqrt{w_{00}^{-1}}, \mathfrak{A}\right).$$

Set also $\delta = \frac{1}{\lambda\sqrt{rd}}$ and introduce the discrete grid $D_\delta = \{\alpha = k\delta : k \in \mathbb{N}, |\alpha| \leq \rho\}$ in the interval $[-\rho, \rho]$. Let $\nu_{k,+}$ (respectively $\nu_{k,-}$) be the (random) point from $D_\delta$ closest to $v_k$ from above (respectively from below), i.e. $\nu_{k,-} \leq v_k \leq \nu_{k,+}$ and $|\nu_{k,\pm} - v_k| \leq \delta$. Denote by $D(v)$ the collection of random vectors $\nu$ of the form $(1, \nu_1, \ldots, \nu_d)^*$, where $\nu_k$ coincides either with $\nu_{k,+}$ of with $\nu_{k,-}$, $k = 1, \ldots, d$. Then, obviously,

$$\max_{\nu \in D(v)} |\nu^* M_T| \geq |v^* M_T|. \tag{2.6}$$

We show now that for every $\nu \in D(v)$, it holds on $\mathfrak{A}$:

$$w_{00}^{-1} \leq \nu^* V \nu \leq (1 + \lambda^{-2}) w_{00}^{-1}. \tag{2.7}$$

Let $\nu \in D(V)$. Then the vector $\Delta = \nu - v = (0, \nu_1 - v_1, \ldots, \nu_d - v_d)^*$ fulfills $\|\Delta\|^2 \leq d\delta^2$. Recall now that $W = V^{-1}$ and $(w_{00}, w_{01}, \ldots, w_{0d})$ is the first row of the matrix $W$, that is,

$$v^* V = w_{00}^{-1}(w_{00}, w_{01}, \ldots, w_{0d}) V = w_{00}^{-1}(1, 0, \ldots, 0).$$

Hence $v^* V v = w_{00}^{-1}$, $v^* V \Delta = \Delta^* V v = 0$, $v^* V v = w_{00}^{-1}$ and

$$\nu^* V \nu = (v + \Delta)^* V (v + \Delta) = w_{00}^{-1} + \Delta^* V \Delta.$$

Since $\Delta^* V \Delta \geq 0$, we get $\nu^* V \nu \geq w_{00}^{-1}$. Moreover, on $\mathfrak{A}$

$$w_{00} \Delta^* V \Delta \leq w_{00} \|V\| \, \|\Delta\|^2 \leq rd\delta^2$$

and (2.7) follows in view of the definition of $\delta$.

Next, being restricted to the set $\mathfrak{A}$, the variable $w_{00}$ fulfills $b \leq w_{00}^{-1} \leq bS$, so that on $\mathfrak{A}$, we get for every $\nu \in D(v)$

$$b \leq \nu^* V \nu \leq (1 + \lambda^{-2}) bS. \tag{2.8}$$

Now (2.6) and (2.7) imply

$$\left\{|v^*M_T| > \lambda\sqrt{w_{00}^{-1}},\, \mathfrak{A}\right\} \;\subseteq\; \bigcup_{\nu\in D(v)} \left\{|\nu^*M_T| > \lambda\sqrt{(1+\lambda^{-2})^{-1}\,\nu^*V\nu}\,,\, \mathfrak{A}\right\},$$

and the use of (2.8) with $\mathfrak{A}_\alpha = \left\{b \le \alpha^*V\alpha \le (1+\lambda^{-2})bS\right\}$ provides

$$\begin{aligned}\left\{|v^*M_T| > \lambda\sqrt{w_{00}^{-1}},\, \mathfrak{A}\right\} &\subseteq \bigcup_{\nu\in D(v)} \left\{|\nu^*M_T| > \lambda\sqrt{(1+\lambda^{-2})^{-1}\,\nu^*V\nu}\,,\, \mathfrak{A}_\alpha\right\} \\ &\subseteq \bigcup_{\alpha\in D_\delta} \left\{|\alpha^*M_T| > \lambda\sqrt{(1+\lambda^{-2})^{-1}\,\alpha^*V\alpha}\,,\, \mathfrak{A}_\alpha\right\}.\end{aligned}$$

Therefore,

$$\boldsymbol{P}\left(|v^*M_T| > \lambda\sqrt{w_{00}^{-1}}\,,\, A\right) \le \sum_{\alpha\in D_\delta} \boldsymbol{P}\left(|\alpha^*M_T| > \lambda\sqrt{(1+\lambda^2)^{-1}\alpha^*V\alpha}\,,\, \mathfrak{A}_\alpha\right).$$

For every $\alpha \in D_\delta$, the process $\alpha^*M_t$ is the scalar square integrable martingale with $\langle\alpha^*M\rangle_T = \alpha^*V\alpha$. Then the application of Theorem 2.1 provides

$$\begin{aligned}&\boldsymbol{P}\left(|\alpha^*M_T| > \lambda\sqrt{(1+\lambda^{-2})^{-1}\alpha^*V\alpha}\,,\, \mathfrak{A}_\alpha\right) \\ &\quad\le 4\left(1+\log S(1+\lambda^{-2})\right)\frac{\lambda}{\sqrt{1+\lambda^{-2}}}\exp\left(-\frac{\lambda^2}{2(1+\lambda^{-2})}+\frac12\right).\end{aligned}$$

Since the number of different elements in $D_\delta$ is at most $(1+2\rho\delta^{-1})^d$, we conclude

$$\begin{aligned}&\boldsymbol{P}\left(|v^*M_T| > \lambda\sqrt{w_{00}^{-1}}, \mathfrak{A}\right) \\ &\quad\le 4\left(1+2\rho\delta^{-1}\right)^d\left(1+\log S(1+\lambda^{-2})\right)\lambda\exp\left(-\frac{\lambda^2}{2(1+\lambda^{-2})}+\frac12\right).\end{aligned}$$

Substituting here $\delta^{-1} = \sqrt{rd}\,\lambda$ and using $\frac{\lambda^2}{1+\lambda^{-2}} \ge \lambda^2 - 1$ for $\lambda^{-2} \le 1/2$, we derive

$$\begin{aligned}\boldsymbol{P}\left(|v^*M_T| > \lambda\sqrt{w_{00}^{-1}}\right) &\le 4e\Big(1+\log(3S/2)\Big)\left(1+2\rho\sqrt{rd}\,\lambda\right)^d\lambda e^{-\frac{\lambda^2}{2}} \\ &\le 4e\log(4S)\left(1+2\rho\sqrt{rd}\,\lambda\right)^d\lambda e^{-\frac{\lambda^2}{2}}\end{aligned}$$

as required. □

2.5. **Coordinate free form.** In the previous section we state the bound for the probability from (1.6) for the special vector $z = (1, 0, \dots, 0)^*$. Here we consider the general

case when $z$ is an arbitrary vector from $\mathbb{R}^{d+1}$ with $\|z\| = 1$. Set

$$\mathfrak{A}_z = \left\{ \begin{array}{c} b \le \dfrac{1}{z^* \langle M \rangle_T^{-1} z} \le bS, \\ \\ z^* \langle M \rangle_T^{-1} z \, \|\langle M \rangle_T\| \le r, \\ \\ \sup\limits_{y \in \mathbb{R}^{d+1} : |y|=1} \dfrac{|y^* \langle M \rangle_T^{-1} z|}{z^* \langle M \rangle_T^{-1} z} \le \rho, \end{array} \right\}.$$

**Theorem 2.3.** *Let $T$ be fixed or stopping time. Then, for every positive constants $b > 0$, $S \ge 1$, $\rho > 0$, $r \ge 1$, and $\lambda \ge \sqrt{2}$*

$$\boldsymbol{P}\left( |z^* \langle M \rangle_T^{-1} M_T| > \lambda \sqrt{z^* \langle M \rangle_T^{-1} z}, \, \mathfrak{A}_z \right) \le 4e \log(4S) \left( 1 + 2\rho\sqrt{rd}\lambda \right)^d \lambda e^{-\frac{\lambda^2}{2}}.$$

*Proof.* For $z = (1, 0, \ldots, 0)^*$, the statement holds by Theorem 2.2. The general case can be reduced to that one simply by changing the coordinate system in the way that $z$ becomes the first coordinate vector. □

2.6. **The ergodic case.** Assume the increments of the martingale $M$ form an ergodic process in a sense that

$$\boldsymbol{P} - \lim_{T \to \infty} \frac{\langle M \rangle_T}{T} = \overline{V}, \tag{2.9}$$

where $\overline{V}$ is a nonsingular deterministic matrix. Denote by $\overline{W} = (\overline{w}_{ij}, \, i, j = 0, \ldots, d)$ the inverse of $\overline{V}$. The ergodic property implies that, for sufficiently large $T$, the random matrix $T \langle M \rangle_T^{-1}$ falls outside any small open vicinity of the limit matrix $\overline{W}$ with a very small probability. This particularly yields that for large $T$ the probability of the event

$$\mathfrak{A}_T = \left\{ \begin{array}{c} \dfrac{1}{2\overline{w}_{00}} \le \dfrac{T}{w_{00}} \le \dfrac{2}{\overline{w}_{00}}, \\ \\ w_{00} \|\langle M \rangle_T\| \le 2\overline{w}_{00} \|\overline{V}\|, \\ \\ \max\limits_{k=1,\ldots,d} \dfrac{|w_{0k}|}{w_{00}} \le 2 \max\limits_{k=1,\ldots,d} \dfrac{|\overline{w}_{0k}|}{\overline{w}_{00}} \end{array} \right\}$$

is closed to 1 and therefore $\boldsymbol{P}(\mathfrak{A}_T^c) = 1 - \boldsymbol{P}(\mathfrak{A}_T)$ is small. In this case, the following result can be useful.

**Proposition 2.1.** *Assume (2.9) with the nonsingular matrix $\overline{V}$. Then there exist constants $C_1$ and $C_2$, depending on $\overline{V}$ only, such that for all $\lambda \ge \sqrt{2}$*

$$\boldsymbol{P}\left( |z^* \langle M \rangle_T^{-1} M_T| > \lambda \sqrt{z^* \langle M \rangle_T^{-1} z} \right) \le C_1 (1 + C_2 \lambda)^d \lambda e^{-\frac{\lambda^2}{2}} + \boldsymbol{P}(\mathfrak{A}_T^c).$$

## 3. Statistical applications

We revert now to the statistical examples from Section 1. First we consider the discrete time model which generalizes Example 1.1. Assume we observe a process $Y_t$, $t \in \mathbb{N}$, and $\mathcal{F}_t$ denotes the $\sigma$-field generated by the observations $Y_s$ with $s \le t$. We also suppose that the observations $Y_t$ follow the equation

$$Y_t = f_t^* \theta + \sigma_t \varepsilon_t, \qquad t = 1, \dots, T, \tag{3.1}$$

where the errors $\varepsilon_t$ are independent standard normal random variables and $f_t$ (resp. $\sigma_t$) is a $\mathbb{R}^p$-valued (resp. $\mathbb{R}_+$-valued) predictable process w.r.t. the filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$, that is, $f_t$ and $\sigma_t$ are completely determined by the observations $Y_1, \dots, Y_{t-1}$. We additionally assume that

$$\boldsymbol{E}\sigma_t^{-2}|f_t|^2 < \infty, \qquad \forall t.$$

Note that the autoregressive model, see Example 1.1, is a particular case of (3.1) with $f_t = (Y_{t-1}, \dots, Y_{t-p})^*$. Similarly to that case, the MLE estimate of the unknown parameter $\theta \in \mathbb{R}^p$ from the observations $Y_t$, $t \le T$, for the model (3.1) reads as follows:

$$\widehat{\theta} = \left( \sum_{t=1}^{T} \sigma_t^{-2} f_t f_t^* \right)^{-1} \sum_{t=1}^{T} \sigma_t^{-2} f_t Y_t$$

and it holds for the estimation error

$$\widehat{\theta} - \theta = \left( \sum_{t=1}^{T} \sigma_t^{-2} f_t f_t^* \right)^{-1} \sum_{t=1}^{T} \sigma_t^{-1} f_t \varepsilon_t = \langle M \rangle_T^{-1} M_T, \tag{3.2}$$

where

$$M_t = \sum_{s=1}^{t} \sigma_s^{-1} f_s \varepsilon_s \qquad \text{and} \qquad \langle M \rangle_t = \sum_{s=1}^{t} \sigma_s^{-2} f_s f_s^*. \tag{3.3}$$

It is straightforward to check that $(M_t,\, t \in \mathbb{N})$ is a square integrable martingale with conditionally Gaussian increments and $(\langle M \rangle_t,\, t \in \mathbb{N})$ is its predictable quadratic variation.

The second application corresponds to the continuous time linear diffusion model (1.7) from Example 1.2.

In the statement below, we treat both models (3.1) and (1.7) simultaneously. Let $T$ be a stopping time w.r.t. the filtration $(\mathcal{F}_t)$ and $\widehat{\theta}$ be the MLE of the unknown parameter $\theta$ from the observations $Y_t$, $t \le T$. Let then $\langle M \rangle_T$ be from (1.9) or (3.3). Define $V = \langle M \rangle_T$ and let $W$ stand for the inverse of $V$. By $w_{k,k'}$ we denote the elements of the matrix $W = V^{-1}$, $k, k' = 1, \dots, p$.

We formulate the result concerning the first coordinate $\widehat{\theta}_1 - \theta_1$ of the vector $\widehat{\theta} - \theta$. The other components of this vector can be treated in a similar way. The assertion is the direct application of Theorem 2.2.

**Theorem 3.1.** *Let $\widehat{\theta}$ be the maximum likelihood estimate of the parameter $\theta$ from observations $Y_t$, $t \le T$, for the model (3.1) (resp. for the model (1.7)) due to (3.2) (resp. (1.8)). For positive constants $b > 0$, $S \ge 1$ $\rho > 0$ and $r \ge 1$, introduce the event*

$$\mathfrak{A} = \left\{ \begin{array}{c} b \le w_{11}^{-1} \le bS, \\ \\ w_{11}\|V\| \le r, \\ \\ |w_{1k}/w_{11}| \le \rho, \quad \forall k = 2, \dots, p \end{array} \right\}.$$

*Then, with any positive $\lambda \ge \sqrt{2}$, it holds*

$$\boldsymbol{P}\left(|\widehat{\theta}_1 - \theta_1| > \lambda\sqrt{w_{11}}, \mathfrak{A}\right) \le 4e\log(4S)\left(1 + 2\rho\sqrt{r(p-1)}\,\lambda\right)^{p-1} \lambda e^{-\frac{\lambda^2}{2}}.$$

## References


[1] Basawa, I.V. and Brockwell, P.J. (1984). Asymptotic conditional inference for regular nonergodic models with an application to autoregressive processes. *Ann. Statist.* **12** 161–171.

[2] Basawa, I.V. and Scott, D.J. (1983). *Asymptotic Optimal Inference for Non-ergodic Models.* Springer New York.

[3] Chan, N.H. and Wei, C.Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16** 367–401.

[4] Cox, D.D. and Llatas, I. (1991). Maximum likelihood type estimation for nearly nonstationarity autoregression time series *Ann. Statist.* **19** 1109–1128.

[5] Grambsch P. (1983). Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. *Ann. Statist.* **11** 68–77.

[6] Härdle,W. and Spokoiny, V. and Tissier (1999). Adaptive estimation for a time inhomogeneous stochastic-volatility model. Unpublished manuscript.

[7] Jacod, J. and Shiryaev, A.N. (1987). *Limit Theorems for Stochastic Processes.* Springer New York.

[8] Jeganathan, P. (1988). On the strong approximation of the distribution of estimators in linear stochastic models, I,II: Stationary and explosive AR models. *Ann. Statist.* **16** no. 3, 1283–1314.

[9] Konev, V.V. and Pergamenshchikov S.S. (1996). On asymptotic minimaxity of fixed accuracy estimators for autoregression parameters. I. Stable process. *Math.-Methods-Statist.* **5** , no. 2, 125–153.

[10] Kreiss, J.P. (1987). On adaptive estimation in stationary ARMA processes. *Ann. Statist.* **15**, 112–133.

[11] Koul, H. and Pflug, G.Ch. (1990). Weakly adaptive estimators in explosive autoregression. *Ann. Statist.* **18** 939–960.

[12] Lai, T.L. and Siegmund, D. (1983). Fixed accuracy estimation of an autoregressive parameter *Ann. Statist.* **11** 478–485.

[13] Liptser, R. and Shiryaev, A.N. (1986). *Theory of Martingales.* Nauka Moscow. (English transl.: Kluwer Acad. Publ. 1989).

[14] Liptser, R. and Spokoiny, V. (1997). On estimating a dynamic function of a stochastic system with averaging. *Ann. Statist.* tentatively accepted. Preprint **381**, Weierstrass Institute, Berlin.

[15] Novikov, A.A. (1972). Sequential estimation of the parameters of processes of diffusion type. *Mat.-Zametki* **12** 627–638.

[16] Rootzen, H. (1983). Central limit theory for martingales via random change of time. in Collection: *Probability and mathematical statistics*. Uppsala Univ., Uppsala, 154–189.

[17] Shiryaev, A. and Spokoiny, V. (1997). On sequential estimation of an autoregressive parameter. *Stochastics and Stoc. Reports*, **60**, 219–240.

[18] White, J.S. (1958). The limiting distribution of the serial correlation coefficient in explosive case. *Ann. Math. Stat.* **29** 1188–1197.

Dept. Electrical Engineering-Systems, Tel Aviv University, 69978 Tel Aviv, Israel

*E-mail address*: liptser@eng.tau.ac.il

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany

*E-mail address*: spokoiny@wias-berlin.de