Bernoulli **11**(4), 2005, 643–664

On the rate of convergence of the maximum likelihood estimator in Brownian semimartingale models

HARRY VAN ZANTEN

Department of Mathematics, Faculty of Sciences, Vrije Universiteit Amsterdam, De Boelelaan1081a, 1081 HV Amsterdam, The Netherlands. E-mail: harry@cs.vu.nl

In this paper we present a unified approach to obtaining rates of convergence for the maximum likelihood estimator (MLE) in Brownian semimartingale models of the form

$$dX_t = \beta_t^{n,\theta}\, dt + \sigma_t^n\, dW_t, \qquad t \leqslant T_n.$$

We show that the rate of the MLE is determined by (an appropriate version of) the entropy of the parameter space with respect to the random metric h_n, defined by

$$h_n^2(\theta, \psi) = \int_0^{T_n} \left(\frac{\beta_s^{n,\theta} - \beta_s^{n,\psi}}{\sigma_s^n} \right)^2 ds.$$

Several known results for the rates in certain popular sub-models of the Brownian semimartingale model are shown to be special cases in our general framework.

Keywords: continuous semimartingale; entropy; exponential inequalities; maximum likelihood estimation; rate of convergence

1. Introduction

In the last few decades, the development of empirical process methods has significantly improved our understanding of the asymptotic behaviour of statistical procedures. An important example is the mathematical description of the intuitive fact that the degree of difficulty of an estimation problem depends on the size, or rather the complexity, of the model. Using the notion of entropy and tools such as uniform exponential inequalities, loose statements of this type can now be made very precise.

For maximum likelihood estimation, there are results for various models which state that the rate of convergence of the maximum likelihood estimator (MLE) is determined by the entropy of the (possibly infinite-dimensional) parameter space relative to the Hellinger metric. Wong and Shen (1995) and van de Geer (1995a) consider independent and identically distributed (i.i.d.) observations from a density p_0 belonging to a set $\mathcal{P}$ of densities with respect to a dominating measure μ. The Hellinger metric on $\mathcal{P}$ is defined by

1350–7265

$$h^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2 \, \mathrm{d}\mu.$$

Denoting the MLE based on the first n observations by $\hat{p}_n$, it turns out that the rate at which $h(\hat{p}_n, p_0)$ vanishes is determined by the bracketing entropy of the parameter space $\mathcal{P}$ with respect to the Hellinger distance (see, for example, Corollary 3.5 of van de Geer 1995a). A similar result is true if we observe a counting process on some time interval $[0, T]$, with a continuous compensator of the form

$$A_t = \int_0^t a_s \, \mu(\mathrm{d}s),$$

for a (possibly random) intensity a belonging to a set $\mathcal{A}$ of intensity processes with respect to a given (possibly random) dominating measure μ. For this model the rate of the MLE is determined by the entropy with bracketing of $\mathcal{A}$ relative to the Hellinger metric, which is defined by

$$h^2(a, b) = \int_0^T (\sqrt{a} - \sqrt{b})^2 \, \mathrm{d}\mu$$

(see van de Geer 1995b, Theorem 4.3). The metric is random in this case, which requires a more careful definition of the appropriate version of entropy than in the i.i.d. situation.

The aim of the present paper is to address, at the same level of abstraction, the problem of finding the rate of the MLE in model

$$\mathrm{d}X_t = \beta_t^{n,\theta} \, \mathrm{d}t + \sigma_t^n \, \mathrm{d}W_t, \qquad t \leqslant T_n. \tag{1.1}$$

Here θ is a parameter which belongs to some abstract parameter space Θ, W is a Brownian motion, $\beta^{n,\theta}$ and σ^n are arbitrary adapted processes such that the stochastic differential equation (1.1) makes sense, and T_n is a non-random number. More precisely, we wish to show that there exists a (random) metric h_n on the parameter space Θ such that an appropriate version of the entropy of Θ with respect to h_n determines the rate. Up till now, results of this type have only been available for certain special cases of (1.1). We mention Nishiyama (1999) who treats the classical signal in Gaussian white noise model, the results of Nishiyama (2000) for the perturbed dynamical system, and van Zanten (2003a) who deals with the ergodic diffusion model. Our main goal is to unify all these results. We define a version of entropy (without bracketing) relative to a random metric and we show that it is the random distance h_n defined by

$$h_n^2(\theta, \psi) = \int_0^{T_n} \left(\frac{\beta_s^{n,\theta} - \beta_s^{n,\psi}}{\sigma_s^n} \right)^2 \mathrm{d}s \tag{1.2}$$

which determines the rate of the MLE in the general model (1.1). This complements the cited results of van de Geer (1995a; 1995b) and Wong and Shen (1995) for i.i.d. data and point processes.

Next, we explain how results for special cases of (1.1) follow from the general theory. Roughly speaking, we show that we can obtain rates for a concrete model if, with large probability, we have a control over the random metric (1.2) of the form

$$\underline{d} \lesssim \frac{h_n}{c_n} \lesssim \bar{d},^{1} \tag{1.3}$$

where the numbers c_n and the metrics $\underline{d}$ and $\bar{d}$ are deterministic. In this case it is the ordinary metric entropy with respect to $\bar{d}$ of a $\underline{d}$-ball around the true parameter which yields the rate of the MLE. Arguing like this, the results of Nishiyama (1999; 2000) and van Zanten (2003a) cited above are easily seen to be special cases in our general framework. For certain null recurrent or transient diffusion models we also have a control of the form (1.3), and hence such models can also be handled by our methods. We illustrate this by considering a transient diffusion model studied, for instance, in Section 3.5 of Kutoyants (2004).

Although our results are all stated for a one-dimensional model (1.1), this restriction is certainly not essential. Generalizations to higher dimensions are straightforward, but omitted for the sake of readability. A non-trivial restriction of the presented results should also be noted. There exist examples of models of the form (1.1) for which we do not have a deterministic control such as (1.3) over the random metric h_n. In such cases it is sometimes possible to obtain random rates of convergence for the MLE. We refer to Loukianova and Loukianov (2003a) for this approach.

The remainder of the paper is organized as follows. In the next section we first provide a general result on rates of convergence of M-estimators, tailored to our purposes. Then in Section 3 we derive a new uniform exponential inequality for families of continuous local martingales, which is an essential ingredient for our main results. Its proof relies on a chaining argument for random metrics, which is given in the Appendix. Section 4 contains the main results of the paper. We first prove that it is the entropy relative to the random metric (1.2) which determines the rate of convergence of the MLE in the model (1.1). We then show that if we have deterministic control like (1.3) over the random metric, then the entropy with respect to the random metric can be replaced by ordinary entropy relative to a deterministic metric. In Section 5 we recover several known results for special cases of the model (1.1) from our general theory.

2. Rates of convergence of M-estimators

In this section we state a result on general M-estimation, which is a straightforward adaptation of well-known results in this area (see, for example, van der Vaart and Wellner 1996; van de Geer 2000). The main reason for the adaptation is that we wish to work with a random metric on the parameter space. Moreover, in the applications we will encounter we can typically only control the metric and the associated entropy on some event which has large probability. Results available in the literature, such as Theorem 3.4.1 of van der Vaart and Wellner (1996), are not directly suited to this situation. The following theorem provides us with sufficient flexibility for our purposes.

[1]We write $a \lesssim b$ if $a \leqslant Cb$ for some positive constant C which is universal, or at least constant throughout the paper.

Theorem 2.1. *Let $\mathbb{Z}$ and Z be random maps on a set Θ, and let θ_0 be a (possibly random) element of Θ. Let $\theta \mapsto \rho(\theta, \theta_0)$ be a random map from Θ to $[0, \infty)$ and let $0 \leqslant \eta \leqslant \infty$ and $s > 0$ be arbitrary. Suppose that for an event A and for all $\delta \in [0, \eta)$ and $x \geqslant 0$,*

$$\sup_{\rho(\theta,\theta_0)\leqslant\eta} Z(\theta) - Z(\theta_0) \leqslant -\rho^s(\theta, \theta_0) \tag{2.1}$$

and

$$\mathbb{P}\left(\sup_{\rho(\theta,\theta_0)\leqslant\delta} ((\mathbb{Z} - Z)(\theta) - (\mathbb{Z} - Z)(\theta_0)) \geqslant x,\ A\right) \leqslant \mathrm{e}^{-x^2/\varphi^2(\delta)}, \tag{2.2}$$

where $\delta \mapsto \varphi(\delta)/\delta^p$ is decreasing on $[0, \eta)$ for some $p < s$. Assume that the numbers $a, r > 0$ satisfy the relation

$$r^s\varphi\left(\frac{a}{r}\right) \leqslant a^q$$

for some $q < s$. Then if $\hat{\theta}$ is a random element of Θ such that $\mathbb{Z}(\hat{\theta}) \geqslant \mathbb{Z}(\theta_0)$, we have that

$$\mathbb{P}(r\rho(\hat{\theta}, \theta_0) > a,\ A) \leqslant C(a, p, q, s)\mathrm{e}^{-a^{2(s-q)}} + \mathbb{P}(\rho(\hat{\theta}, \theta_0) > \eta,\ A),$$

where $C(a, p, q, s) < \infty$ is a constant with the property that $C(a, p, q) \downarrow 0$ as $a \to \infty$.

Proof. To simplify the notation, set $\mathbb{G}(\theta) = (\mathbb{Z} - Z)(\theta) - (\mathbb{Z} - Z)(\theta_0)$. For $j = 1, 2, \ldots,$ define the random sets $S_j = \{\theta \in \Theta : a2^{j-1} < r\rho(\theta, \theta_0) \leqslant a2^j\}$. Then we have

$$\mathbb{P}(r\rho(\hat{\theta}, \theta_0) > a,\ A) \leqslant \sum_{j:a2^j\leqslant r\eta} \mathbb{P}(\hat{\theta} \in S_j,\ A) + \mathbb{P}(\rho(\hat{\theta}, \theta_0) > \eta,\ A).$$

If $\hat{\theta} \in S_j$, the supremum of the map $\mathbb{Z} - \mathbb{Z}(\theta_0)$ over S_j is non-negative, so

$$\mathbb{P}(\hat{\theta} \in S_j,\ A) \leqslant \mathbb{P}\left(\sup_{\theta\in S_j} \mathbb{Z}(\theta) - \mathbb{Z}(\theta_0) \geqslant 0,\ A\right).$$

For $\theta \in S_j$ and $a2^j \leqslant r\eta$ we have $\rho(\theta, \theta_0) \leqslant \eta$. So by assumption (2.1) we have for every j appearing in the sum, for some constant $c > 0$,

$$\begin{aligned}\mathbb{P}(\hat{\theta} \in S_j,\ A) &\leqslant \mathbb{P}\left(\sup_{\theta\in S_j} \mathbb{G}(\theta_0) \geqslant c\rho^s(\theta, \theta_0),\ A\right)\\ &\leqslant \mathbb{P}\left(\sup_{r\rho(\theta,\theta_0)\leqslant a2^j} \mathbb{G}(\theta_0) \geqslant c\left(\frac{a2^{j-1}}{r}\right)^s,\ A\right).\end{aligned}$$

By assumption (2.2), it follows that the last probability is bounded, up to a constant, by

$$\exp\left(-\frac{c^2a^{2s}2^{s(2j-2)}}{r^{2s}\varphi^2(a2^j/r)}\right).$$

Since $\delta \mapsto \varphi(\delta)/\delta^p$ is decreasing on $[0, \eta)$, we have $\varphi(a2^j/r) \leqslant 2^{pj}\varphi(a/r)$ for every j

appearing in the sum. Using also the assumption on r we obtain, for every j appearing in the sum,

$$\mathbb{P}(\hat{\theta} \in S_j, A) \leqslant \mathrm{e}^{-Da^{2(s-q)}2^{2j(s-p)}},$$

for a constant $D > 0$ depending only on s. This implies that the assertion of the theorem holds true with the constant

$$C(a, p, q, s) = \sum_{j \geqslant 1} \mathrm{e}^{-a^{2(s-q)}(D4^{(s-p)j}-1)},$$

which clearly has the desired properties. □

In the next section we will apply Theorem 2.1 to random maps $\mathbb{Z}$ and Z which have the property that $\mathbb{Z}(\theta) - Z(\theta)$ is the terminal point of a continuous local martingale which depends on θ. The choice of random distance ρ will be such that (2.1) is satisfied automatically (with $s = 2$). To verify (2.2) we need an appropriate uniform maximal inequality for continuous martingales. This is the subject of the next section.

3. A uniform exponential inequality for continuous martingales

The uniform exponential inequality we derive in this section can be viewed as a generalization of the maximal inequality for continuous local martingales of Nishiyama (1999). Nishiyama's result deals with a collection $\mathcal{M}$ of continuous martingales, metricized by a given non-random metric d. It gives an entropy bound for expectations of quantities of the form

$$\sup_{\substack{M,N \in \mathcal{M} \\ d(M,N) \leqslant \delta}} |M_t - N_t|.$$

Our new result essentially gives entropy bounds for expectations of quantities such as

$$\sup_{\substack{M,N \in \mathcal{M} \\ \langle M-N \rangle_t \leqslant \delta}} |M_t - N_t|,$$

where $\langle M \rangle$ is the quadratic variation process, or bracket, of the continuous martingale M. So instead of considering a given deterministic metric on $\mathcal{M}$, we endow the class of martingales with the natural random metric induced by the brackets. We use the same version of 'entropy relative to a random distance' as, for instance, van de Geer (2002).

The classical Bernstein inequality for continuous local martingales says that if M is a continuous local martingale vanishing at 0, with quadratic variation process $\langle M \rangle$, then for all $x, L > 0$,

$$\mathbb{P}\left(\sup_{t \geqslant 0} |M_t| \geqslant x, \langle M \rangle_\infty \leqslant L\right) \leqslant e^{-x^2/2L} \tag{3.1}$$

(see for instance Revuz and Yor 1999, pp. 153–154). Now suppose we have a collection $\mathcal{M}$

of continuous local martingales, defined on a single filtered probability space. We endow $\mathcal{M}$ with the random semimetric $\rho_{\mathcal{M}}$, defined by $\rho^2_{\mathcal{M}}(M, M') = \langle M - M' \rangle_\infty$. Using this notation, (3.1) states that for $M, M' \in \mathcal{M}$,

$$\mathbb{P}\big(\|M - M'\|_\infty \geqslant x,\, \rho^2_{\mathcal{M}}(M, M') \leqslant L\big) \leqslant e^{-x^2/2L} \tag{3.2}$$

for all $x, L > 0$.

It is convenient to express this inequality in terms of Orlicz norms. Recall that for a Young function ψ (an increasing, convex function on $\mathbb{R}_+$ with $\psi(0) = 0$), the ψ-norm of a random variable X is defined as

$$\|X\|_\psi = \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|X|}{C}\right) \leqslant 1\right\}.$$

A sub-Gaussian inequality like (3.2) can be formulated in terms of the ψ_2-norm, where $\psi_2(x) = \exp(x^2) - 1$. If a random variable X has a distribution with tails satisfying $\mathbb{P}(|X| > x) \leqslant K\exp(-Cx^2)$, then $\|X\|^2_{\psi_2} \leqslant (1 + K)/C$ (van der Vaart and Wellner 1996, Lemma 2.2.1). Hence, (3.2) translates into

$$\|\|M - M'\|_\infty 1_{\{\rho_{\mathcal{M}} M, M') \leqslant L\}}\|_{\psi_2} \leqslant 2L \tag{3.3}$$

for $M, M' \in \mathcal{M}$ and $L > 0$. Conversely, it is also true that a bound on the ψ_2-norm as in (3.3) leads to a sub-Gaussian tail bound like (3.2). Hence, the two formulations are equivalent. We use the Orlicz norms because they are more convenient from a technical point of view.

Below we present a uniform extension of inequality (3.3). We will keep M' fixed, and let M range over the entire class $\mathcal{M}$. An upper bound will be given in terms of the 'size' of $\mathcal{M}$ with respect to the random distance $\rho_{\mathcal{M}}$. To measure this size we use the notion of 'entropy with respect to a random distance' or 'partioning entropy' (without bracketing) as in van de Geer (2002); see also van de Geer (1995b) for a version with bracketing.

For the general definition, consider a collection $\mathcal{X}$ of random elements of an arbitrary set (usually a vector space), defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Suppose that for every pair $X, Y \in \mathcal{X}$ we have a non-negative random variable $\rho(X, Y)$, and that these have the property that, almost surely, $\rho(X, Y) \leqslant \rho(X, Z) + \rho(Z, Y)$ for all $X, Y, Z \in \mathcal{X}$.

Definition 3.1. *For every event $A \in \mathcal{F}$ and $0 < \varepsilon \leqslant \delta \leqslant \infty$ we define the covering number $N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A)$ as the smallest number n for which there exist $X_1, \ldots, X_n \in \mathcal{X}$ such that, for every $X \in \mathcal{X}$, there exists an index $i \in \{1, \ldots, n\}$ such that $\rho(X, X_i) \leqslant \varepsilon$ on the event $A \cap \{\rho(X, Y) \leqslant \delta\}$. For $\delta = \infty$ we write $N(\varepsilon, \mathcal{X}, Y, \infty, \rho, A) = N(\varepsilon, \mathcal{X}, \rho, A)$ and $N(\varepsilon, \mathcal{X}, \rho, \Omega)$ is abbreviated to $N(\varepsilon, \mathcal{X}, \rho)$.*

Let us emphasize that in this definition it is essential that the map $\mathcal{X} \to \{X_1, \ldots, X_n\}$ which assigns to $X \in \mathcal{X}$ an X_i such that $\rho(X, X_i) \leqslant \varepsilon$ on the event $A \cap \{\rho(X, Y) \leqslant \delta\}$, is deterministic. It may depend on anything else however; in particular, it will typically depend on the event A. The number $N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A)$ should be thought of as the ε-covering number of the ball around Y of ρ-radius δ.

In concrete cases the event A may be used to control the random distance ρ by a deterministic metric, in order to obtain bounds in terms of the covering numbers for a suitable deterministic distance, for which many useful results exist in the literature. The following simple lemma is useful in this regard.

Lemma 3.1. *For $Y \in \mathcal{X}$, suppose there exist constants $c, C > 0$ and deterministic pseudo-metrics $\underline{d}$ and $\bar{d}$ on $\mathcal{X}$, such that, on the event A, $c\underline{d}(X, Y) \leqslant \rho(X, Y)$ for all $X \in \mathcal{X}$ and $\rho(X, X') \leqslant C\bar{d}(X, X')$ for all $X, X' \in \mathcal{X}$. Then $N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A) \leqslant N(\varepsilon/C, \mathcal{X}_{\delta/c}, \bar{d})$, where $\mathcal{X}_\delta = \{X \in \mathcal{X}\colon \underline{d}(X, Y) \leqslant \delta\}$.*

Proof. Suppose that $\underline{d}(X, Y) > \delta/c$. Then on A we have $\rho(X, Y) > \delta$, so that $A \cap \{\rho(X, Y) \leqslant \delta\} = \varnothing$. It follows that we only have to consider $X \in \mathcal{X}_{\delta/c}$. For $N = N(\varepsilon/C, \mathcal{X}_{\delta/c}, \bar{d})$, let $\{X_1, \ldots, X_N\}$ be an ε/C-net of $\mathcal{X}_{\delta/c}$ for the metric $\bar{d}$. Then for $X \in \mathcal{X}_{\delta/c}$ there exists an X_i such that on $A \cap \{\rho(X, Y) \leqslant \delta\}$ we have $\rho(X, X_i) \leqslant C\bar{d}(X, X_i) \leqslant \varepsilon$. □

The covering numbers of Definition 3.1 have precisely the properties needed to make inequality (3.3) uniform in M, using a straightforward chaining method. It is shown in the Appendix that this works in great generality. Consider a class $\mathcal{X}$ of random elements of a vector space $\mathcal{V}$, a random semimetric ρ on $\mathcal{X}$ and a Young function ψ such that

$$\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$$

for some constant $c > 0$. Then if, for the event A,

$$\| \|X - Y\|_{\mathcal{V}} 1_{A\cap\{\rho(X,Y)\leqslant\delta\}} \|_\psi \leqslant \delta$$

for $X, Y \in \mathcal{X}$ and $\delta > 0$, we have the uniform inequality

$$\left\| \sup_{X\in\mathcal{X}} \|X - Y\|_{\mathcal{V}} 1_{A\cap\{\rho(X,Y)\leqslant\delta\}} \right\|_\psi \leqslant C\int_0^\delta \psi^{-1}(N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A))\,\mathrm{d}\varepsilon,^2$$

where C is positive constant that only depends on ψ, and ψ^{-1} is the inverse of ψ (see Theorem A.1).

We now apply this general result to the class of continuous local martingales $\mathcal{M}$, endowed with the random distance $\rho_{\mathcal{M}}$. Note that for the Young function $\psi_2(x) = \exp(x^2) - 1$ we have $\psi_2^{-1}(x) = \sqrt{\log(1 + x)}$. For $x \geqslant 2$ it holds that $\log(1 + x) \leqslant \log(2x) = \log 2 + \log x \leqslant 2\log x$, so $\psi^{-1}(x) \lesssim \sqrt{\log x}$. Hence, in view of (3.3), we arrive at the following result.

Theorem 3.2. *Let $\mathcal{M}$ be a collection of continuous local martingales, and for $M, M' \in \mathcal{M}$ define $\rho^2_{\mathcal{M}}(M, M') = \langle M - M'\rangle_\infty$. Then for every $M' \in \mathcal{M}$,*

[2]Expectations of possibly non-measurable suprema should be interpreted as outer expectations.

$$\left\| \sup_{M \in \mathcal{M}} \sup_{t \geq 0} |M_t - M'_t| 1_{A \cap \{\langle M - M' \rangle_\infty \leq \delta^2\}} \right\|_{\psi_2} \lesssim \int_0^\delta \sqrt{\log N(\varepsilon, \mathcal{M}, M', \delta, \rho_{\mathcal{M}}, A)} \, \mathrm{d}\varepsilon,$$

for every event A and all $\delta > 0$.

By stopping the martingales it is easily seen that the result of the theorem is also true with ∞ replaced by any stopping time τ. Indeed, we simply apply the preceding theorem to the class $\mathcal{M}' = \{M^\tau : M \in \mathcal{M}\}$ of stopped martingales. (We use the standard notation $M_t^\tau = M_{\tau \wedge t}$.) We have that $\langle M^\tau - M'^\tau \rangle_\infty = \langle M - M' \rangle_\tau$, leading to the following equivalent result.

Theorem 3.3. *Let $\mathcal{M}$ be a collection of continuous local martingales and τ a stopping time. For $M, M' \in \mathcal{M}$ define $\rho^2_{\mathcal{M}}(M, M') = \langle M - M' \rangle_\tau$. Then for every $M' \in \mathcal{M}$,*

$$\left\| \sup_{M \in \mathcal{M}} \sup_{t \leq \tau} |M_t - M'_t| 1_{A \cap \{\langle M - M' \rangle_\tau \leq \delta^2\}} \right\|_{\psi_2} \lesssim \int_0^\delta \sqrt{\log N(\varepsilon, \mathcal{M}, M', \delta, \rho_{\mathcal{M}}, A)} \, \mathrm{d}\varepsilon,$$

for every event A and all $\delta > 0$.

A bound on the ψ_2-norm implies a sub-Gaussian bound on tail probabilities. Indeed, by definition of the Orlicz norm and Markov's inequality,

$$\mathbb{P}(|X| \geq x) \leq 1 \wedge \frac{1}{\psi_2(x/\|X\|_{\psi_2})} \leq 2\mathrm{e}^{-x^2/\|X\|^2_{\psi_2}}$$

for all $x > 0$. So in terms of tail probabilities, Theorem 3.3 reads as follows.

Theorem 3.4. *Let $\mathcal{M}$ be a collection of continuous local martingales and τ a stopping time. For $M, M' \in \mathcal{M}$ define $\rho^2_{\mathcal{M}}(M, M') = \langle M - M' \rangle_\tau$. Then for every $M' \in \mathcal{M}$, any event A and all $x, \delta > 0$,*

$$\mathbb{P}\left(\sup_{M \in \mathcal{M}} \sup_{t \leq \tau} |M_t - M'_t| 1_{A \cap \{\langle M - M' \rangle_\tau \leq \delta^2\}} \geq x \right) \leq 2\mathrm{e}^{-x^2/C\varphi^2(\delta)},$$

where

$$\varphi(\delta) = \int_0^\delta \sqrt{\log N(\varepsilon, \mathcal{M}, M', \delta, \rho_{\mathcal{M}}, A)} \, \mathrm{d}\varepsilon$$

and $C > 0$ is a universal constant.

We remark that Theorem 3.3 extends the maximal inequality presented by Nishiyama (1999). Nishiyama considered a class $\mathcal{M}$ of continuous local martingales endowed with a non-random metric d. He introduced the so-called 'quadratic modulus', which is defined as

$$\|\mathcal{M}\|_{d,\tau} = \sup_{d(M,N)>0} \frac{\sqrt{\langle M - N \rangle_\tau}}{d(M, N)}.$$

If we use this quantity to control the random metric $\rho_{\mathcal{M}}$, it is not very hard to obtain Theorem 2.3 of Nishiyama (1999) as a corollary of our Theorem 3.3. So Theorem 3.3 can be viewed as an extension of Nishiyama's maximal inequality, giving a uniform exponential inequality for continuous martingales without referring to some auxiliary deterministic metric. Instead, the size of the class $\mathcal{M}$ is measured by the natural distance induced by the brackets of the martingales.

4. The rate of the MLE in the Brownian semimartingale model

For every $n \in \mathbb{N}$, let $(\Omega^n, \mathcal{F}^n, (\mathcal{F}^n_t), \mathbb{P}^n)$ be a filtered probability space. On this stochastic basis, suppose that we have a standard Brownian motion W^n and adapted processes X^n, β^{n,θ_0} and σ^n satisfying

$$X^n_t = X^n_0 + \int_0^t \beta^{n,\theta_0}_s \, \mathrm{d}s + \int_0^t \sigma^n_s \, \mathrm{d}W^n_s, \qquad t \leqslant T_n, \tag{4.1}$$

where T_n is a positive (non-random) number. It is implicitly understood that the processes β^{n,θ_0} and σ^n are such that the Lebesgue and Itô integrals are well defined.

We suppose that θ_0 is an unknown element of the abstract parameter space Θ. To estimate it we use the MLE defined by

$$\hat{\theta}_n = \operatorname*{arg\,max}_{\theta \in \Theta} \int_0^{T_n} \frac{\beta^{n,\theta}_s}{(\sigma^n_s)^2} \, \mathrm{d}X^n_s - \frac{1}{2} \int_0^{T_n} \left(\frac{\beta^{n,\theta}_s}{\sigma^n_s} \right)^2 \mathrm{d}s. \tag{4.2}$$

The MLE is assumed to exist with probability one.

Under certain regularity conditions the expression on the right-hand side of (4.2) is precisely the log-likelihood $\log \mathrm{d}\mathbb{P}^n_\theta / \mathrm{d}\mathbb{P}^n_0$, where $\mathbb{P}^n_\theta$ is the law of the process X^n with θ instead of θ_0 in (4.1) (observed up to time T_n) and $\mathbb{P}^n_0$ is the law of the process X^n which satisfies (4.1) with $\beta \equiv 0$; (see, for example, Liptser and Shiryayev (1977) or Jacod and Shiryaev (1987). However, we will not need to impose precise conditions implying absolute continuity. For our purposes it suffices to assume the minimal requirement that for every $\theta \in \Theta$,

$$\int_0^{T_n} \left(\frac{\beta^{n,\theta}_s}{\sigma^n_s} \right)^2 \mathrm{d}s < \infty \tag{4.3}$$

$\mathbb{P}^n$-almost surely, which is needed to ensure that the integrals in (4.2) are well defined.

It is easily seen that this set-up fits into the setting of Theorem 2.1. Indeed, define the random maps $\mathbb{Z}_n$ and Z_n on Θ by

$$\mathbb{Z}_n(\theta) = \int_0^{T_n} \frac{\beta^{n,\theta}_s}{(\sigma^n_s)^2} \, \mathrm{d}X^n_s - \frac{1}{2} \int_0^{T_n} \left(\frac{\beta^{n,\theta}_s}{\sigma^n_s} \right)^2 \mathrm{d}s$$

and

$$Z_n(\theta) = \int_0^{T_n} \frac{\beta_s^{n,\theta}\beta_s^{n,\theta_0}}{(\sigma_s^n)^2}\,\mathrm{d}s - \frac{1}{2}\int_0^{T_n}\left(\frac{\beta_s^{n,\theta}}{\sigma_s^n}\right)^2 \mathrm{d}s.$$

Then

$$Z_n(\theta) - Z_n(\theta_0) = -\frac{1}{2}h_n^2(\theta, \theta_0),$$

where h_n is the random semimetric on Θ defined by

$$h_n^2(\theta, \psi) = \int_0^{T_n}\left(\frac{\beta_s^{n,\theta} - \beta_s^{n,\psi}}{\sigma_s^n}\right)^2 \mathrm{d}s. \tag{4.4}$$

If we let h_n play the role of ρ in Theorem 2.1, condition (2.1) is automatically fulfilled (for $s = 2$ and $\eta = \infty$). As for condition (2.2), observe that

$$(\mathbb{Z}_n - Z_n)(\theta) - (\mathbb{Z}_n - Z_n)(\theta_0) = \int_0^{T_n} \frac{\beta_s^{n,\theta} - \beta_s^{n,\theta_0}}{\sigma_s^n}\,\mathrm{d}W_s^n.$$

Hence, the required uniform exponential inequality is provided by Theorem 3.4.

A straightforward application of Theorems 2.1 and 3.4 now yields the following result, which states that it is the entropy with respect to the metric (4.4) which determines the rate of the MLE. This complements the analogous results of van de Geer (1995a) and Wong and Shen (1995) for i.i.d. observations, and van de Geer (1995b) for counting processes.

Theorem 4.1. *Let $n \in \mathbb{N}$ be fixed and let A be an arbitrary event. Suppose that*

$$\int_0^{\delta} \sqrt{\log N(\varepsilon, \Theta, \theta_0, \delta, h_n, A)}\,\mathrm{d}\varepsilon \leqslant \varphi(\delta),$$

for all $\delta \in [0, \eta)$, where $\eta \leqslant \infty$ and φ is a function such that $\delta \mapsto \varphi(\delta)/\delta^p$ is decreasing on the interval $[0, \eta)$ for some $p < 2$. Moreover, assume that the numbers $a, r > 0$ satisfy the relation

$$r^2\varphi\left(\frac{a}{r}\right) \leqslant a^q$$

for some $q < 2$. Then

$$\mathbb{P}^n(rh_n(\hat{\theta}, \theta_0) > a, A) \leqslant C(a, p, q)\mathrm{e}^{-a^{4-2q}} + \mathbb{P}^n(h_n(\hat{\theta}, \theta_0) > \eta, A),$$

where $C(a, p, q) < \infty$ is a constant with the property that $C(a, p, q)\downarrow 0$ as $a \to \infty$.

In specific Brownian semimartingale models the random metric (4.4) typically converges, after a suitable normalization, to a deterministic metric on the parameter space. The latter can be viewed as the 'natural' distance for that specific model. In particular, the normalized random metric will usually be equivalent to some non-random metric, with large probability. If we have such deterministic control over the metric (4.4), the rate of convergence of the MLE is determined by the entropy of the parameter space relative to this natural non-random metric.

Theorem 4.2. *Suppose we have semimetrics* $\underline{d}$ *and* $\bar{d}$ *on* Θ *such that*

$$\int_0^\delta \sqrt{\log N(\varepsilon, \underline{\Theta}_\delta, \bar{d})}\, \mathrm{d}\varepsilon \lesssim \varphi(\delta)$$

for all $\delta \in [0, \eta)$*, where* $\eta \leqslant \infty$ *and* φ *is a function such that* $\delta \mapsto \varphi(\delta)/\delta^p$ *is decreasing on the interval* $[0, \eta)$ *for some* $p < 2$*, with* $\underline{\Theta}_\delta = \{\theta \in \Theta : \underline{d}(\theta, \theta_0) \leqslant \delta\}$*. Moreover, let* r_n *and* c_n *be sequences converging to infinity, satisfying*

$$r_n^2 \varphi\left(\frac{1}{r_n}\right) \lesssim c_n.$$

Finally, suppose that for every $\gamma > 0$ *there exists an event* A *with probability at least* $1 - \gamma$ *and constants* $c, C > 0$ *such that:*

(i) *there exists an* $a_0 > 0$ *such that on the event* A*,*

$$c\underline{d}(\theta, \theta_0) \leqslant \frac{1}{c_n} h_n(\theta, \theta_0)$$

for all $n \in \mathbb{N}$ *and* $\theta \in \Theta$ *with* $a_0/r_n < \underline{d}(\theta, \theta_0) < \eta$*;*

(ii) *on the event* A*,*

$$\frac{1}{c_n} h_n(\theta, \psi) \leqslant C\bar{d}(\theta, \psi)$$

for all $n \in \mathbb{N}$ *and* $\theta, \psi \in \Theta$.

Then if $\bar{d}(\hat{\theta}_n, \theta_0)$ *converges to* 0 *in probability,* $\underline{d}(\hat{\theta}_n, \theta_0) = O_P(r_n^{-1})$*. If the conditions involving* η *are satisfied for* $\eta = \infty$*, the assumption of consistency can be dropped.*

Proof. Let $\gamma > 0$ be given and consider the event A and the numbers $a_0, c, C > 0$ given in assumptions (i) and (ii). We fix n for a moment and introduce the new parameter space $\widetilde{\Theta} = \{\theta \in \Theta : \eta > \underline{d}(\theta, \theta_0) > a_0/r_n\} \cup \{\theta_0\}$. By (i) and (ii) and Lemma 3.1 we have

$$N(\varepsilon, \widetilde{\Theta}, \theta_0, \delta, h_n, A) \leqslant N\left(\frac{\varepsilon}{c_n C}, \widetilde{\Theta}_{\delta/c}, \bar{d}\right) = N\left(\frac{c\varepsilon}{c_n C}, \widetilde{\Theta}_\delta, \bar{d}\right),$$

where $\widetilde{\Theta}_\delta = \{\theta \in \widetilde{\Theta}: \underline{d}(\theta, \theta_0) \leqslant \delta\}$. It follows that for $\delta < c_n C\eta/c$,

$$\int_0^\delta \sqrt{\log N(\varepsilon, \widetilde{\Theta}, \theta_0, \delta, h_n, A)}\, \mathrm{d}\varepsilon \leqslant \frac{c_n C}{c} \int_0^{c\delta/c_n C} \sqrt{\log N(\varepsilon, \Theta_\delta, \bar{d})}\, \mathrm{d}\varepsilon$$

$$= \frac{c_n C}{c} \varphi\left(\frac{c\delta}{c_n C}\right) =: \varphi_n(\delta).$$

On the event $\{\eta > \underline{d}(\hat{\theta}_n, \theta_0) > a/r_n\}$ for $a > a_0$, the MLE does not change if we replace the parameter space Θ by the smaller space $\widetilde{\Theta}$. By assumption (i) we have for $a \geqslant a_0$ and n large enough,

$$\mathbb{P}^n(\eta > \underline{d}(\hat{\theta}_n, \theta_0) > a/r_n, A) \leqslant \mathbb{P}^n\Big(r_n h_n(\hat{\theta}_n, \theta_0) > ac_n c, A\Big).$$

Note that the assumption on φ implies that for $a \geqslant 1$ and n large enough,

$$\varphi\left(\frac{a}{r_n}\right) \leqslant \varphi\left(\frac{1}{r_n}\right) a^p.$$

Hence, the numbers $r_n' = r_n/c_n c$ and the function φ_n satisfy

$$(r_n')^2 \varphi_n\left(\frac{a}{r_n'}\right) \leqslant a^p$$

By Theorem 4.1 (applied with $\widetilde{\Theta}$ instead of Θ, φ_n instead φ, $\eta_n = \eta c_n C/c$ instead of η and r_n' instead of r_n) and assumption (ii), it follows that

$$\mathbb{P}^n(r_n \underline{d}(\hat{\theta}_n, \theta_0) > a, A) \leqslant C(a, p, p)\mathrm{e}^{-a^{4-2p}} + \mathbb{P}^n(c\bar{d}(\hat{\theta}_n, \theta_0) > \eta, A) + \mathbb{P}^n(\underline{d}(\hat{\theta}_n, \theta_0) \geqslant \eta).$$

By assumption the second and third terms on the right-hand side vanish as $n \to \infty$ (or vanish identically if $\eta = \infty$). Since A has probability at least $1 - \gamma$, it follows that for a large enough,

$$\limsup_{n\to\infty} \mathbb{P}^n(r_n \underline{d}(\hat{\theta}_n, \theta_0) > a) \leqslant 2\gamma.$$

This completes the proof. □

Applications of the theorem to some specific models are considered in the next section. Let us remark here that condition (i) of the theorem can sometimes be verified by showing that

$$\sup_{\underline{d}(\theta,\theta_0)<\eta} \left| \frac{1}{c_n} h_n(\theta, \theta_0) - \underline{d}(\theta, \theta_0) \right| = O_P\left(\frac{1}{r_n}\right). \tag{4.5}$$

Indeed, if this holds there exists a constant $L > 0$ such that with probability at least $1 - \gamma$,

$$r_n \left| \frac{1}{c_n} h_n(\theta, \theta_0) - \underline{d}(\theta, \theta_0) \right| \leqslant L$$

for all $n \in \mathbb{N}$ and $\theta \in \Theta$ such that $\underline{d}(\theta, \theta_0) < \eta$. It follows that if we choose a_0 so large that $L/a_0 \leqslant 1/2$, then on this event we have

$$\frac{1}{c_n} h_n(\theta, \theta_0) \geqslant \frac{1}{2} \underline{d}(\theta, \theta_0)$$

for all $n \in \mathbb{N}$ and θ such that $a_0/r_n < \underline{d}(\theta, \theta_0) < \eta$. This is precisely the requirement of condition (i) of the theorem.

The restriction on the rates r_n implied by (4.5) is in fact quite natural. We should view the rate at which $h_n/c_n \to \underline{d}$ as the 'parametric' rate for that specific model. The restriction is then simply that the rate r_n of the MLE over an arbitrary, possibly infinite-dimensional parameter space should not be faster than the parametric rate.

The pseudo-metric $\bar{d}$ can in many examples be taken equal to the natural metric $\underline{d}$. In order to verify condition (ii) it may be convenient to take a larger distance in certain cases. However, it should be noted that this could lead to a less tight upper bound for the rate of the MLE.

5. Examples

The aim of this last section is to show that our results for the general model (4.1) allow us to recover some well-known results for special cases. For a specific model one has to consider the random metric (4.4) and find the deterministic metric with which, after a suitable normalization, it is equivalent with large probability. Candidates for this normalization and natural non-random metric are usually easily found in concrete cases. The asymptotic properties of the particular model can then be used to verify conditions (i) and (ii) of Theorem 4.2. This yields an upper bound for the rate of convergence of the MLE relative to the natural metric for the model under consideration.

Since our interest in this paper is in rates of convergence, we will assume existence and consistency of the MLE. For results on these matters, see for instance Kutoyants (2004), Loukianova and Loukianov (2003b), van Zanten (2001) and the references therein.

5.1. Signal in white noise

The first example is the 'signal in white noise model', given by the stochastic differential equation

$$\mathrm{d}X_t^n = \theta_0(t)\,\mathrm{d}t + \sigma_n\,\mathrm{d}W_t, \qquad t \leqslant T.$$

Here the time horizon T is fixed, θ_0 is an unknown function in $\Theta \subseteq L^2[0, T]$, called the signal, and the number $\sigma_n > 0$ is the noise level. It is assumed that $\sigma_n \to 0$ as $n \to \infty$. This model has been studied in detail with the help of entropy methods by Nishiyama (1999). In this subsection we briefly show that it fits into our general framework.

In this case the metric (4.4) is in fact deterministic itself, and is given by

$$h_n(\theta, \psi) = \frac{1}{\sigma_n}\|\theta - \psi\|_{L^2[0,T]}.$$

Hence, in Theorem 4.2 we take $c_n = 1/\sigma_n$ and for $\underline{d}$ and $\bar{d}$ we take the $L^2[0, T]$ distance. The theorem then yields the following result for this model, first obtained by Nishiyama (1999).

Proposition 5.1. *Suppose that*

$$\int_0^\delta \sqrt{\log N(\varepsilon, \Theta_\delta, \|\cdot\|_{L^2[0,T]})}\,\mathrm{d}\varepsilon \leqslant \varphi(\delta)$$

for all $\delta \in [0, \eta)$*, where* $\eta \leqslant \infty$ *and* φ *is a function such that* $\delta \mapsto \varphi(\delta)/\delta^p$ *is decreasing on* *the interval* $[0, \eta)$ *for some* $p < 2$, *and* $\Theta_\delta = \{\theta \in \Theta : \|\theta - \theta_0\|_{L^2[0,T]} \leqslant \delta\}$. *Let* r_n *be sequence converging to infinity such that*

$$r_n^2 \varphi\left(\frac{1}{r_n}\right) \lesssim \frac{1}{\sigma_n}.$$

Then if the MLE $\hat{\theta}_n$ *exists and is consistent,*

$$\|\hat{\theta}_n - \theta_0\|_{L^2[0,T]} = O_P\left(\frac{1}{r_n}\right)$$

as $n \to \infty$. *If the conditions are satisfied for* $\eta = \infty$, *the assumption of consistency is not necessary.*

For applications of Proposition 5.1 to various concrete examples of Θ, such as smooth parametric classes or classes of monotone functions, see Nishiyama (1999; 2000).

5.2. Perturbed dynamical system

Next we consider the model

$$\mathrm{d}X_t^n = \theta_0(X_t^n)\,\mathrm{d}t + \sigma_n\,\mathrm{d}W_t, \qquad t \leqslant T,\ X_0^n = x_0,$$

where the true parameter θ_0 belongs to some class of functions Θ, and the noise level σ_n is a positive number that vanishes as $n \to \infty$. The random semimetric h_n is now given by

$$h_n(\theta, \psi) = \frac{1}{\sigma_n}\sqrt{\int_0^T (\theta(X_t^n) - \psi(X_t^n))^2\,\mathrm{d}t}.$$

To obtain the candidate for the 'natural' metric $\underline{d}$, observe that it is plausible that as $n \to \infty$, the process X^n will tend to the solution $t \mapsto x_t$ of the unperturbed ordinary differential equation $\mathrm{d}x_t = \theta_0(x_t)\,\mathrm{d}t$. In particular, we can expect that

$$\sigma_n^2 h_n^2(\theta, \theta_0) \to \int_0^T (\theta(x_t) - \theta_0(x_t))^2\,\mathrm{d}t =: \underline{d}^2(\theta, \theta_0).$$

If we consider, for instance, a parameter space Θ such that the functions in the space are uniformly Lipschitz continuous in the sense that

$$\sup_{\theta \in \Theta} \sup_{x \neq y} \frac{|\theta(x) - \theta(y)|}{|x - y|} < \infty, \tag{5.1}$$

the argument above can be made precise and we can apply Theorem 4.2. The proof of the following proposition uses some ideas from Section 6.4 of Nishiyama (2000).

Proposition 5.2. *Suppose that* (5.1) *holds for the parameter space* Θ *and* $\sup_{\theta \in \Theta}\|\theta\|_\infty < \infty$. *Moreover, suppose that*

$$\int_0^\delta \sqrt{\log N(\varepsilon, \Theta_\delta, \|\cdot\|_\infty)}\,\mathrm{d}\varepsilon \lesssim \varphi(\delta)$$

for all $\delta \in [0, \eta)$, *where* $\eta \leqslant \infty$ *and* φ *is a function such that* $\delta \mapsto \varphi(\delta)/\delta^p$ *is decreasing on the interval* $[0, \eta)$ *for some* $p < 2$, *and* $\Theta_\delta = \{\theta \in \Theta : \underline{d}(\theta, \theta_0) \leqslant \delta\}$. *Let* r_n *be a sequence converging to infinity such that* $\sigma_n r_n$ *remains bounded and*

$$r_n^2 \varphi\left(\frac{1}{r_n}\right) \lesssim \frac{1}{\sigma_n}.$$

Then if the MLE $\hat{\theta}_n$ *exists and* $\|\hat{\theta}_n - \theta_0\|_\infty \to 0$ *in (outer) probability,*

$$\underline{d}(\hat{\theta}_n - \theta_0) = O_P\left(\frac{1}{r_n}\right)$$

as $n \to \infty$. *If the conditions are satisfied for* $\eta = \infty$, *the assumption of consistency is not necessary.*

Proof. Observe that if θ_0 is Lipschitz, the Gronwall inequality (Karatzas and Shreve 1991, pp. 287–288) implies that

$$\sup_{t \leqslant T} |X_t^n - x_t| = O_P(\sigma_n)$$

as $n \to \infty$. By (5.1), it follows that

$$\sup_{\theta \in \Theta} \sup_{t \leqslant T} |\theta(X_t^n) - \theta(x_t)| = O_P(\sigma_n)$$

as $n \to \infty$. Hence, we have

$$\begin{aligned}\left|\sigma_n h_n(\theta, \psi) - \underline{d}(\theta, \psi)\right|^2 &\leqslant \int_0^T \Big((\theta(X_t^n) - \theta(x_t)) - (\psi(X_t^n) - \psi(x_t))\Big)^2 \mathrm{d}t \\ &\leqslant 2\int_0^T \big(\theta(X_t^n) - \theta(x_t)\big)^2 \mathrm{d}t + 2\int_0^T \big(\psi(X_t^n) - \psi(x_t)\big)^2 \mathrm{d}t,\end{aligned}$$

whence

$$\sup_{\theta, \psi \in \Theta} \left|\sigma_n h_n(\theta, \psi) - \underline{d}(\theta, \psi)\right| = O_P(\sigma_n).$$

This shows that condition (i) of Theorem 4.2 is satisfied with $c_n = 1/\sigma_n$, provided that the rate r_n is not faster than $1/\sigma_n$, which is the parametric rate for this model (see the remarks following the theorem). Clearly, condition (ii) is satisfied with $\bar{d}$ the uniform distance $\|\cdot\|_\infty$. □

Using the preceding proposition it is straightforward to recover, for instance, the result of Nishiyama (2000, pp. 111–112) dealing with a class Θ such that all functions $\theta \in \Theta$ vanish outside some bounded set $I \subseteq \mathbb{R}$ and $\Theta \subseteq C_M^\alpha(I)$, $C_M^\alpha(I)$ being the space of functions f on I such that $\|f\|_\alpha \leqslant M$, where

$$\|f\|_\alpha = \max_{k \leqslant \underline{\alpha}} \|f^{(k)}\|_\infty + \sup_{x,y} \frac{|f^{(\underline{\alpha})}(x) - f^{(\underline{\alpha})}(y)|}{|x-y|^{\alpha-\underline{\alpha}}}$$

and $\underline{\alpha}$ is the greatest integer (strictly) smaller than α. Using a well-known entropy bound for this function space an upper bound,

$$r_n = \sigma_n^{-2\alpha/(2\alpha+1)},$$

for the rate of the MLE with respect to the natural metric $\underline{d}$ can be obtained.

5.3. Ergodic diffusions

In this subsection we consider the stochastic differential equation

$$\mathrm{d}X_t = \theta_0(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \qquad t \leqslant T_n.$$

Under certain regularity conditions (see Karatzas and Shreve 1991, Section 5.5), this equation generates a strong Markov process on a (possibly unbounded) open interval $I \subseteq \mathbb{R}$, with scale function s_0 given by

$$s_0(x) = \int_{x_0}^{x} \exp\left(-2\int_{x_0}^{y} \frac{\theta_0(z)}{\sigma^2(z)}\right)\mathrm{d}y$$

(x_0 is an arbitrary, but fixed point in the state space) and speed measure

$$m_0(\mathrm{d}x) = \frac{\mathrm{d}x}{s_0'(x)\sigma^2(x)}.$$

We assume that m_0 has finite total mass, that is, $m_0(I) < \infty$. Then the diffusion is ergodic, and the normalized speed measure $\mu_0 = m_0/m_0(I)$ is the unique invariant probability measure. For simplicity, the initial law of the diffusion is supposed to be degenerate in some point $x \in I$. The endpoint T_n of the observation interval is assumed to tend to infinity as $n \to \infty$.

In this model the semimetric h_n in (4.4) is given by

$$h_n^2(\theta, \psi) = \int_0^{T_n} \left(\frac{\theta(X_t) - \psi(X_t)}{\sigma(X_t)}\right)^2 \mathrm{d}t.$$

To define the metric $\underline{d}$ we choose a fixed compact $J \subseteq I$ and define

$$\underline{d}(\theta, \psi) = \sqrt{\int_J \left(\frac{\theta - \psi}{\sigma}\right)^2 d\mu_0} = \left\|\frac{\theta - \psi}{\sigma} 1_J\right\|_{L^2(\mu_0)}.$$

For $\bar{d}$ we take

$$\bar{d}(\theta, \psi) = \left\|\frac{\theta - \psi}{\sigma}\right\|_{L^2(\mu_0)}.$$

The fact that these metrics satisfy conditions (i) and (ii) of Theorem 4.2 follows from results of van Zanten (2003b). To see this, let $(l_t(x),\ t \geqslant 0,\ x \in I)$ be the diffusion local time of the process X relative to the speed measure, see, for instance, Itô and McKean (1965). Then

$$\frac{1}{T_n} h_n^2(\theta, \psi) = \frac{m_0(I)}{T_n} \int_I \left(\frac{\theta(x) - \psi(x)}{\sigma(x)} \right)^2 l_{T_n}(x)\, \mu_0(\mathrm{d}x)$$

$$\geqslant\ m_0(I) \inf_{x \in J} \frac{1}{T_n} l_{T_n}(x) \underline{d}^2(\theta, \psi).$$

According to Theorem 3.2 of van Zanten (2003b) it holds that

$$\sup_{x \in J} \left| \frac{1}{T_n} l_{T_n}(x) - \frac{1}{m_0(I)} \right| \to 0$$

in probability. In particular, we have

$$\inf_{x \in J} \frac{1}{T_n} l_{T_n}(x) \geqslant \frac{1}{2 m_0(I)}$$

with probability tending to 1, which covers condition (i). Observe that we have $c_n = \sqrt{T_n}$ in this case. Next, note that

$$\frac{h_n^2(\theta, \psi)}{c_n^2} \leqslant \left(\frac{1}{T_n} \sup_{x \in I} l_{T_n}(x) \right) m_0(I) \bar{d}^2(\theta, \psi).$$

This shows that condition (ii) is satisfied, since the local time has the property that

$$\frac{1}{T_n} \sup_{x \in I} l_{T_n}(x) = O_P(1)$$

as $n \to \infty$ (see van Zanten (2003b), Theorem 3.1).

Summarizing, our general theorem yields the following result for the ergodic diffusion model.

Proposition 5.3. *Let $J \subseteq I$ be compact. Suppose that*

$$\int_0^\delta \sqrt{\log N\left(\varepsilon,\ \Theta_\delta,\ \left\| \frac{\cdot}{\sigma} \right\|_{L^2(\mu_0)} \right)}\, \mathrm{d}\varepsilon \lesssim \varphi(\delta)$$

for all $\delta \in [0, \eta)$, where $\eta \leqslant \infty$ and φ is a function such that $\delta \mapsto \varphi(\delta)/\delta^p$ is decreasing on the interval $[0, \eta)$ for some $p < 2$, and

$$\Theta_\delta = \left\{ \theta \in \Theta\colon \left\| \frac{\theta - \theta_0}{\sigma} 1_J \right\|_{L^2(\mu_0)} \leqslant \delta \right\}.$$

Let r_n be a sequence converging to infinity such that

$$r_n^2 \varphi\left(\frac{1}{r_n} \right)\ \lesssim\ \sqrt{T_n}.$$

Then if the MLE $\hat{\theta}_n$ exists and $\|(\hat{\theta}_n - \theta_0)/\sigma\|_{L^2(\mu_0)} \to 0$ in probability,

$$\left\| \frac{\hat{\theta}_n{}' - \theta_0}{\sigma} 1_J \right\|_{L^2(\mu_0)} = O_P\left(\frac{1}{r_n}\right).$$

If the conditions are satisfied for $\eta = \infty$, the assumption of consistency is not necessary.

The main difference with Theorem 4 of van Zanten (2003a) is that for the preceding result we do not need the somewhat artificial condition on the tails of the invariant law μ_0 needed in the latter paper. The new result is therefore applicable to a wider class of ergodic diffusion models. The entropy calculations in concrete cases are very similar, and we refer to van Zanten (2003a) for examples.

5.4. Null recurrent and transient diffusions

As mentioned in the Introduction, we can also handle null recurrent or transient diffusion models for which we have deterministic control over the random metric (4.4). As an illustration, consider the model

$$\mathrm{d}X_t = \theta_0 |X_t|^\kappa \,\mathrm{d}t + \sigma \,\mathrm{d}W_t, \qquad X_0 = x_0,\ t \leqslant T_n,$$

where θ_0 is an unknown element of an open set $\Theta \subseteq (\alpha, \beta)$ for certain $\alpha, \beta > 0$, $\sigma > 0$, $\kappa \in (0, 1)$ and $T_n \to \infty$ as $n \to \infty$. This model is studied, for instance, in Section 3.5.2 of Kutoyants (2004). The MLE $\hat{\theta}_n$ for this model exists, is consistent, and is asymptotically normal with rate $T_n^{(1+\kappa)/(2-2\kappa)}$ (Kuyoyants 2004, Proposition 3.45). This rate can easily be recovered using our general results.

Indeed, the metric h_n is in this case given by

$$h_n^2(\theta, \psi) = \left(\frac{\theta - \psi}{\sigma}\right)^2 \int_0^{T_n} |X_t|^{2\kappa} \,\mathrm{d}t.$$

It is well known that for the present model, $X_t \sim C t^{1/(1-\kappa)}$ almost surely as $t \to \infty$, for a positive constant C (see, for example, Gikhman and Skorohod 1969). This implies, in particular, that

$$\frac{1}{T_n^{(1+\kappa)/(1-\kappa)}} \int_0^{T_n} |X_t|^{2\kappa} \,\mathrm{d}t$$

converges almost surely to a positive limit, and hence conditions (i) and (ii) of Theorem 4.2 are fulfilled with $c_n = T_n^{(1+\kappa)/(2-2\kappa)}$ and $\underline{d} = \bar{d}$ the Euclidean distance. The entropy integral is in this case equal to a multiple of δ, which leads to the rate $r_n = c_n = T_n^{(1+\kappa)/(2-2\kappa)}$, as desired.

We remark that similar null recurrent or transient diffusion models, such as the null recurrent model studied by Höpfner and Kutoyants (2003), the transient Ornstein–Uhlenbeck process and the extension considered by Dietz and Kutoyants (2003), can be handled in the same way. The crucial point is that in all these cases, the metric h_n

converges at a deterministic rate. When this is not the case our method brakes down, and one has to resort to different techniques. See for instance Loukianova and Loukianov (2003a), who consider random rates of convergence.

Appendix: Entropy inequalities for random metrics

Let $\mathcal{X}$ be an arbitrary collection of random elements of some normed vector space $\mathcal{V}$. The norm of $\mathcal{V}$ is simply denoted by $|\cdot|$. It is well known that if there exists a metric d on $\mathcal{X}$ such that $\|X - Y\|_\psi \leqslant d(X, Y)$ for all $X, Y \in \mathcal{X}$, then a chaining argument can be used to obtain the uniform inequality

$$\left\| \sup_{d(X,Y)\leqslant\delta} |X - Y| \right\|_\psi \lesssim \int_0^\delta \psi^{-1}(N(\varepsilon, \mathcal{X}_\delta, d))\, \mathrm{d}\varepsilon. \tag{A.1}$$

Here $\mathcal{X}_\delta$ is the ball of d-radius δ around Y and $N(\varepsilon, \mathcal{X}_\delta, d)$ is the minimal number of balls of radius ε that are needed to cover $\mathcal{X}_\delta$. (See Chapter 11 of Ledoux and Talagrand (1991), or Chapter 2.2 of van der Vaart and Wellner (1996) for maximal inequalities of this form.)

Here we derive the extension of this result to collections of random variables endowed with a random pseudo-metric. We assume that for all $X, Y \in \mathcal{X}$ we have a non-negative random variable $\rho(X, Y)$, and that these have the property that almost surely $\rho(X, Y) \leqslant \rho(X, Z) + \rho(Z, Y)$ for all $X, Y, Z \in \mathcal{X}$. Recall Definition 3.1 of the covering numbers with respect to the random norm. We assume throughout that the Young function ψ has the property that

$$\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$$

for some constant c.

Theorem A.1. *Let $\mathcal{X}$ be a collection of random elements of a normed vector space $\mathcal{V}$. Suppose that $\mathcal{X}$ is endowed with a random pseudo-metric ρ, and let A be an arbitrary event. Suppose that for all $X, Y \in \mathcal{X}$ and $\delta > 0$,*

$$\| |X - Y| 1_{A\cap\{\rho(X,Y)\leqslant\delta\}} \|_\psi \leqslant \delta.$$

Then for every $Y \in \mathcal{X}$ and all $\delta > 0$,

$$\left\| \sup_{X\in\mathcal{X}} |X - Y| 1_{A\cap\{\rho(X,Y)\leqslant\delta\}} \right\|_\psi \leqslant C \int_0^\delta \psi^{-1}(N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A))\, \mathrm{d}\varepsilon,$$

where $C > 0$ is a constant that only depends on ψ.

Proof. Replacing $\mathcal{X}$ by the collection $\{X - Y \colon X \in \mathcal{X}\}$, we reduce to the case that $Y = 0$, so we can assume $Y = 0$ without loss of generality. For $j = 0, 1, 2, \ldots,$ let $\mathcal{X}_j$ be a set of $n_j = N(\delta 2^{-j}, \mathcal{X}, Y, \delta, \rho, A)$ elements of $\mathcal{X}$ such that, for every $X \in \mathcal{X}$, there exists a fixed $X_j \in \mathcal{X}_j$ such that $\rho(X, X_j) \leqslant \delta 2^{-j}$ on the event $A \cap \{\rho(X, 0) \leqslant \delta\}$. So for every $X \in \mathcal{X}$, we have an infinite 'chain' $X, \ldots, X_j, X_{j-1}, \ldots, X_0$ that starts at X and ends at some element $X_0 \in \mathcal{X}_0$. By construction $X_j 1_{A\cap\{\rho(X,0)\leq\delta\}} \to X 1_{A\cap\{\rho(X,0)\leq\delta\}}$ in ψ-norm as $j \to \infty$, so we have the equality

$$(X - X_0)1_{A\cap\{\rho(X,0)\leq\delta\}} = \sum_{j=0}^{\infty}(X_{j+1} - X_j)1_{A\cap\{\rho(X,0)\leq\delta\}}$$

in ψ-norm. Since on the event $A \cap \{\rho(X, 0) \leq \delta\}$ we have that $\rho(X, X_j) \leq \delta 2^{-j}$ and hence $\rho(X_{j+1}, X_j) \leq \rho(X, X_{j+1}) + \rho(X, X_j) \leq \delta 2^{-j+1}$, it follows that

$$\begin{aligned}|X - X_0|1_{A\cap\{\rho(X,0)\leq\delta\}} &\leq \sum_{j=0}^{\infty} |X_{j+1} - X_j| 1_{A\cap\{\rho(X_{j+1},X_j)\leq\delta 2^{-j+1}\}} \\ &\leq \sum_{j=0}^{\infty} \max|U - V| 1_{A\cap\{\rho(U,V)\leq\delta 2^{-j+1}\}},\end{aligned}$$

where the maximum in the jth term is over all 'links' (U, V) from $\mathcal{X}_{j+1}$ to $\mathcal{X}_j$. There are at most n_j such links, and by assumption the ψ-norm of each random variable appearing in the maximum is bounded by $\delta 2^{-j+1}$. Hence, the ψ-norm of the jth term in the sum is bounded by $C\psi^{-1}(n_j)\delta 2^{-j+1}$, where C is a constant that only depends on ψ (see, for instance, van der Vaart and Wellner 1996, Lemma 2.2.2). It follows that

$$\begin{aligned}\left\|\sup_{X\in\mathcal{X}}|X - X_0|1_{A\cap\{\rho(X,0)\leq\delta\}}\right\|_{\psi} &\leq 2C\sum_{j=0}^{\infty}\psi^{-1}(N(\delta 2^{-j}, \mathcal{X}, Y, \delta, \rho, A))\delta 2^{-j} \\ &\leq 4C\int_0^{\delta}\psi^{-1}(N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A))\,\mathrm{d}\varepsilon. \qquad \text{(A.2)}\end{aligned}$$

The same argument as before implies that

$$|X|1_{A\cap\{\rho(X,0)\leq\delta\}} \leq |X - X_0|1_{A\cap\{\rho(X,0)\leq\delta\}} + |X_0|1_{A\cap\{\rho(X_0,0)\leq 2\delta\}},$$

where X_0 is the endpoint in $\mathcal{X}_0$ of the chain starting at X. Hence, by (A.2) we have

$$\left\|\sup_{X\in\mathcal{X}}|X|1_{A\cap\{\rho(X,0)\leq\delta\}}\right\|_{\psi} \leq 4C\int_0^{\delta}\psi^{-1}(N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A))\,\mathrm{d}\varepsilon + \|\max|X_0|1_{A\cap\{\rho(X_0,0)\leq 2\delta\}}\|_{\psi},$$

where the latter maximum is over all endpoints X_0 in $\mathcal{X}_0$ of chains that start at some $X \in \mathcal{X}$. There are at most n_0 such endpoints, so by the assumption of the theorem and Lemma 2.2.2 of van der Vaart and Wellner (1996) we have

$$\|\max|X_0|1_{A\cap\{\rho(X_0,0)\leq 2\delta\}}\|_{\psi} \leq 2C\delta\psi^{-1}(N(\delta, \mathcal{X}, Y, \delta, \rho, A)).$$

If we combine this with what we already had we obtain

$$\|\sup_{X\in\mathcal{X}}|X|1_{A\cap\{\rho(X,0)\leq\delta\}}\|_{\psi} \leq 4C\int_0^{\delta}\psi^{-1}(N(\varepsilon, \mathcal{X}, Y, \delta, \rho, A))\,\mathrm{d}\varepsilon + 2C\delta\psi^{-1}(N(\delta, \mathcal{X}, Y, \delta, \rho, A)).$$

This yields the statement of the theorem. □

Acknowledgement

Thanks are due to the referees and an associate editor for their valuable suggestions and references.

References

Dietz, H.M. and Kutoyants, Yu, A. (2003) Parameter estimation for some non-recurrent solutions of SDE. *Statist. Decisions*, **21**(1), 29–45.

Gikhman, I.I. and Skorokhod, A.V. (1969) *Introduction to the Theory of Random Processes*. Philadelphia: W.B. Saunders.

Höpfner, R. and Kutoyants, Yu.A. (2003) On a problem of statistical inference in null recurrent diffusions. *Statist. Inference Stochastic Process.* **6**(1), 25–42.

Itô, K. and McKean, H.P., Jr. (1965) *Diffusion Processes and Their Sample Paths*. Berlin, Springer-Verlag.

Jacod, J. and Shiryaev, A.N. (1987) *Limit Theorems for Stochastic Processes*. Berlin: Springer-Verlag.

Karatzas, I. and Shreve, S.E. (1991) *Brownian motion and stochastic calculus*. New York: Springer-Verlag.

Kutoyants, Yu.A. (2004) *Statistical Inference for Ergodic Diffusion Processes*. New York: Springer-Verlag.

Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces*. Berlin: Springer-Verlag.

Liptser, R.S. and Shiryayev, A.N. (1977) *Statistics of Random Processes I*. New York: Springer-Verlag.

Loukianova, D. and Loukianov, O. (2003a) Almost sure rate of convergence of maximum likelihood estimators for multidimensional diffusions. Preprint.

Loukianova, D. and Loukianov, O. (2003b) Uniform law of large numbers and consistency of estimators for multidimensional Harris diffusions. Preprint.

Nishiyama, Y. (1999) A maximal inequality for continuous martingales and *M*-estimation in a Gaussian white noise model. *Ann. Statist.*, **27**, 675–696.

Nishiyama, Y. (2000) *Entropy Methods for Martingales*, CWI Tract 128. Amsterdam: Centrum voor Wiskunde en Informatica.

Revuz, D. and Yor, M. (1999) *Continuous Martingales and Brownian Motion* (3rd edn). Berlin: Springer-Verlag.

van de Geer, S.A. (1995a) The method of sieves and minimum contrast estimators. *Math. Methods Statist.*, **4**(1), 20–38.

van de Geer, S.A. (1995b) Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.*, **23**, 1779–1801.

van de Geer, S.A. (2000) *Applications of Empirical Process Theory*. Cambridge: Cambridge University Press.

van de Geer, S.A. (2002) On Hoeffding's inequality for dependent random variables. In *Empirical Process Techniques for Dependent Data*, pp. 161–169. Boston: Birkhäuser.

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes with Applications to Statistics*. New York: Springer-Verlag.

van Zanten, J.H. (2001) A note on consistent estimation of multivariate parameters in ergodic diffusion models. *Scand. J. Statist.*, **28**, 617–623.

van Zanten, J.H. (2003a) On empirical processes for ergodic diffusions and rates of convergence of *M*-estimators. *Scand. J. Statist.*, **30**, 443–458.

van Zanten, J.H. (2003b) On uniform laws of large numbers for ergodic diffusions and consistency of estimators. *Statist. Inference Stochastic Process.*, **6**(2), 199–213.

Wong, W.H. and Shen, X. (1995) Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, **23**, 339–362.

Received February 2004 and revised November 2004