Autoregression Approximation of a Nonparametric Diffusion Model

Grigori Milstein and Michael Nussbaum
Weierstrass Institute, Berlin

April 1996

Abstract

We consider a model of small diffusion type where the function which governs the drift term varies in a nonparametric set. We investigate discrete versions of this continuous model with respect to statistical equivalence, in the sense of the asymptotic theory of experiments. It is shown that an Euler difference scheme as a discrete version of the stochastic differential equation is asymptotically equivalent in the sense of Le Cam's deficiency distance, when the discretization step decreases with the noise intensity ϵ. We thus obtain a nonparametric version of diffusion limit results for autoregression. It follows that in the continuous diffusion model, discrete sampling on a uniform grid is asymptotically sufficient. The key technical step utilizes the notion of Hellinger process from semimartingale theory.

1 Introduction

Consider the problem of estimating the function f from an observed diffusion process $y(t),\ t \in [0,1]$, which satisfies an Ito stochastic differential equation

$$dy(t) = f(y(t))dt + \epsilon\ dW(t),\ t \in [0,1],\ y(0) = 0 \tag{1}$$

where $dW(t)$ is Gaussian white noise and ϵ is a small parameter. Suppose that the function f belongs to some a priori set Σ, nonparametric in general. Kutoyants (1985) constructed estimates of the function f the squared L_2-risk of which decreases with rate $\epsilon^{4m/(2m+1)}$ if the function f has m bounded derivatives. These are the standard

1991 *Mathematics Subject Classification.* Primary 62 M 10; Secondary 62 G 07, 62 B 15, 60 J 60, 60 H 10.

Key words and phrases. Nonparametric experiments, deficiency distance, likelihood ratio process, stochastic differential equation, autoregression, diffusion sampling, asymptotic sufficiency.

nonparametric rates of convergence which also hold in the problem of 'signal recovery in Gaussian white noise'

$$dx(t) = f(t)dt + \epsilon\, dW(t),\ t \in [0,1]. \tag{2}$$

Brown and Low (1996) found that the continuous model (2) is asymptotically equivalent to its discrete counterpart, i. e. the nonparametric regression

$$x_i = f(t_i) + \epsilon n^{1/2}\xi_i,\ i = 1, \ldots, n \tag{3}$$

with a uniform grid $t_i = (i-1)/n$ and standard normal variables ξ_i, provided that f varies in a nonparametric subset of $L_2(0,1)$ defined by a moderate smoothness condition and n tends to infinity not too slowly. The framework was asymptotic equivalence in the sense of Le Cam's deficiency distance Δ. In this paper we address the analogous question with respect to discretizing the stochastic differential equation model (1).
The discretization to consider is suggested by the theory of numerical solution of stochastic differential equations, see e. g. Kloeden, Platen (1993). There the function f is assumed known and an approximation is sought for the solution y from a discrete difference scheme, where the continuous Gaussian white noise $dW(t)$ is substituted by a sequence of i. i. d. normal random variables. The simplest such difference scheme is the one of Euler type. To derive it, observe that the process y satisfies

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(y(u))du + \epsilon(W(t_{i+1}) - W(t_i)),\ i = 1, \ldots, n. \tag{4}$$

For the approximation, the integral is substituted by $n^{-1}f(y(t_i))$ and $W(t_{i+1}) - W(t_i)$ is written $n^{-1/2}\xi_i$ where $\xi_i,\ i = 1, \ldots, n$ is a sequence of i. i. d. standard normal variables. Introducing an approximate solution process y_i defined on the grid only (y_i corresponding to grid point t_i), one gets a sequence of successive approximations

$$y_{i+1} = y_i + n^{-1}f(y_i) + \epsilon\, n^{-1/2}\xi_i,\ i = 1, \ldots, n,\ y_1 = 0. \tag{5}$$

This directly extends the classical Euler scheme for approximating solutions of deterministic differential equations, with random noise added in each step to mimick the random continuous solution y of (1). It is then shown that the process y_i on the discrete grid approximates the solution y of (1) in some probabilistic sense as $n \to \infty$, if one considers the ξ_i as actually coming from W via $\xi_i = n^{1/2}(W(t_{i+1}) - W(t_i))$.
The statistical task may be seen as the converse: given y satisfying (1), reconstruct f. A natural question which then arises is whether inference may be based on the grid values of the solution process $y(t_i)$ only. But these values satisfying (4) still depend on the whole path of y via the integral over $[t_i, t_{i+1}]$. Going a step further, one might then ask whether estimating f in (1) is equivalent to estimating f from the 'truly' discretized process y_i in (5).
Our strategy for comparing two models is based on the following basic concepts from asymptotic decision theory (Le Cam (1986)). Consider experiments $\mathbb{E} = (\Omega, \mathcal{B}, (P_f,\ f \in \Sigma))$ all having parameter space Σ. If Ω is Polish space and $\mathcal{B}$ is the pertaining Borel sigma-algebra then the measure space $(\Omega, \mathcal{B})$ and the experiment $\mathbb{E}$ are called *polish.*

If the family $(P_f,\ f \in \Sigma)$ is dominated then $\mathbb{E}$ is called *dominated.* Henceforth in this paper all experiments are assumed polish and dominated (p. d.). Suppose we have two p. d. experiments $\mathbb{E}_i = (\Omega_i, \mathcal{B}_i, (P_{i,f},\ f \in \Sigma))$ with the same parameter space Σ, but possibly different sample spaces Ω_i. $\mathbb{E}_2$ is a *randomization* of $\mathbb{E}_1$ if there is a Markov kernel: $K : \Omega_1 \times \mathcal{B}_2 \mapsto [0,1]$ such that $P_{2,f} = P_{1,f},\ f \in \Sigma$. In that case, $\mathbb{E}_1$ is also said to be *more informative* than $\mathbb{E}_2$ ($\mathbb{E}_1 \succeq \mathbb{E}_2$). The p. d. experiments $\mathbb{E}_i$, $i = 1,2$ are *statistically isomorphic* or *equivalent* if simultaneously $\mathbb{E}_1 \succeq \mathbb{E}_2$ and $\mathbb{E}_2 \succeq \mathbb{E}_1$. The equivalence class constitutes an *experiment type.* If T is a sufficient statistic in $\mathbb{E}_1$ with values in $(X, \mathcal{F})$ and if the experiment $\mathbb{E}_1^T = (X, \mathcal{F}, (T \circ P_{1,f},\ f \in \Sigma))$ is polish then $\mathbb{E}_1^T$ is equivalent to $\mathbb{E}_1$. If $\mathbb{E}_i$, $i = 1,2$ are given on the same measurable space $(\Omega_1, \mathcal{B}_1) = (\Omega_2, \mathcal{B}_2)$ then their total variation distance is

$$\Delta_0(\mathbb{E}_1, \mathbb{E}_2) = \sup_{f \in \Sigma} \|P_{1,f} - P_{2,f}\|_{TV}$$

where $\|\cdot\|_{TV}$ is the total variation distance of measures. The *one-sided deficiency* of $\mathbb{E}_1$ with respect to $\mathbb{E}_2$ is

$$\delta(\mathbb{E}_1, \mathbb{E}_2) = \inf(\Delta_0(\mathbb{E}_2, \mathbb{E}_1^*) : \mathbb{E}_2 \succeq \mathbb{E}_1^*)$$

the infimum extending over $\mathbb{E}_1^*$ for which $\Delta_0(\mathbb{E}_2, \mathbb{E}_1^*)$ is defined. We have $\delta(\mathbb{E}_1, \mathbb{E}_2) = 0$ if and only if $\mathbb{E}_1 \succeq \mathbb{E}_2$. Le Cam's *deficiency pseudodistance* between arbirtrary p. d. experiments is

$$\Delta(\mathbb{E}_1, \mathbb{E}_2) = \max(\delta(\mathbb{E}_1, \mathbb{E}_2), \delta(\mathbb{E}_2, \mathbb{E}_1)).$$

The concepts pertainining to sequences $\mathbb{E}_{i,\epsilon} = (\Omega_\epsilon, \mathcal{B}_\epsilon, (P_{i,f,\epsilon},\ f \in \Sigma))$, $\epsilon \to 0$are the following.

Definition. (i) *Two p. d. sequences* $\mathbb{E}_{1,\epsilon}$, $\mathbb{E}_{2,\epsilon}$ *on a common measurable space* $(\Omega_\epsilon, \mathcal{B}_\epsilon)$ *are asymptotically total variation equivalent if*

$$\Delta_0(\mathbb{E}_{1,\epsilon}, \mathbb{E}_{2,\epsilon}) \to 0 \text{ as } \epsilon \to 0.$$

(ii) *For sequences* $\mathbb{E}_{1,\epsilon}$, $\mathbb{E}_{2,\epsilon}$ *as in* (i), *let* T_ϵ *be a sequence of statistics on* $(\Omega_\epsilon, \mathcal{B}_\epsilon)$ *with values in polish measure spaces* $(X_\epsilon, \mathcal{F}_\epsilon)$. *The sequence* T_ϵ *is asymptotically sufficient in* $\mathbb{E}_{1,\epsilon}$ *if* T_ϵ *is sufficient in* $\mathbb{E}_{1,\epsilon}$ *and* $\mathbb{E}_{1,\epsilon}, \mathbb{E}_{2,\epsilon}$ *are asymptotically total variation equivalent.*

(iii) *Two arbitrary p. d. sequences* $\mathbb{E}_{1,\epsilon}, \mathbb{E}_{2,\epsilon}$ *are asymptotically equivalent in the sense of Le Cam if*

$$\Delta(\mathbb{E}_{1,\epsilon}, \mathbb{E}_{2,\epsilon}) \to 0 \text{ as } \epsilon \to 0.$$

Asymptotic total variation equivalence implies asymptotic equivalence in the Le Cam sense. In the latter case, we also call $\mathbb{E}_{1,\epsilon}, \mathbb{E}_{2,\epsilon}$ accompanying sequences; if all $\mathbb{E}_{2,\epsilon}$ are of the same type we recover Le Cam's notion of limit of experiments. Then the type of $\mathbb{E}_{1,\epsilon}$ has a (strong) limit which is the type of $\mathbb{E}_{1,\epsilon}$. Henceforth the term asymptotic equivalence will be used in the sense of (iii).

The motivation for considering Δ comes from decision theory. Indeed, another characterization is as follows. Consider a measurable space of decisions, randomized decison rules r in an experiment $\mathbb{E}$, a nonnegative loss function l bounded by M and define the risk $R_{l,\mathbb{E}}(r, f)$ of a decision rule r for loss l at parameter f in the usual manner. Let $\rho(\mathbb{E})$ be the set of decision rules in $\mathbb{E}$ for the given decision problem. Then we have: for all $r_1 \in \rho(\mathbb{E}_1)$ there is $r_2 \in \rho(\mathbb{E}_2)$ such that

$$\sup_{f\in\Sigma} |R_{l,\mathbb{E}_1}(r_1, f) - R_{l,\mathbb{E}_2}(r_2, f)| \leq M\Delta(\mathbb{E}_1, \mathbb{E}_2)$$

In other words, asymptotic equivalence in the sense of Le Cam means that for any sequence of decision functions in one experiment there exists another sequence in the other experiment which is as good aymptotically, for any sequence of loss functions which is uniformly bounded.
The concept $\Delta(\mathbb{E}_{1,\epsilon}, \mathbb{E}_{2,\epsilon}) \to 0$ in statistical theory has found application mostly for the case of limits, where the sequence $\mathbb{E}_{2,\epsilon}$ is constant (does not depend on ϵ). This has been achieved predominantly for parametric statistical models after localization, which are then locally asymptotically normal (Le Cam (1986)). With the seminal paper of Brown and Low (1996) it turned out however, that for nonparametric models with parameters in function spaces where estimation problems are of the "ill-posed" class, the deficiency distance can also be applied successfully, if one considers accompanying sequences of experiments rather than limits. The first example was the asymptotic equivalence of Gaussian nonparametric regression (3) to signal recovery in Gausian white noise (2). A result for i. i. d. observations with unknown density f is obtained in Nussbaum (1996) (asymptotic equivalence of density estimation and Gaussian white noise). A direct application of asymptotic equivalence to nonparametric estimation is presented in Korostelev and Nussbaum (1996).
The present paper is the first result of this type for a stochastic process model involving dependence, giving a statistical discretization of a diffusion process model with parameter in function space. It turns out that notion the Hellinger process from semimartingale theory is also very useful for approximating nonparametric experiments; for parametric limit theory cp. Höpfner, Jacod, Ladelli (1990). We believe thus that the method has a potentially much wider applicability, for the statistical comparison of discrete and continuous stochastic process models, as well as for accompanying sequences in the spirit of limit theorems.

2 Methodology and Main Result

Accompanying sequences of experiments were utilized in Le Cam (1985) for a parametric model of independent observations and a sequence of Gaussian shift experiments. Later in Le Cam and Yang (1990) a general method was proposed for establishing asymptotic equivalence. Suppressing the index ϵ, suppose $\mathbb{E}_i$ are dominated by a probability measure Π_i on $(\Omega_i, \mathcal{B}_i)$. Consider $\Lambda_i(f)(\omega) = dP_{i,f}/d\Pi_i(\omega)$, $\omega \in \Omega_i$ as a random variable under Π_i and consider the corresponding stochastic process indexed by f. Now find versions of these processes on a common probability space; more precisely, find a polish probability space $(\Omega^*, \mathcal{B}^*, \Pi^*)$ and maps $U_i : (\Omega^*, \mathcal{B}^*, \Pi^*) \mapsto (\Omega_i, \mathcal{B}_i)$ such that $U_i \circ \Pi^* = \Pi_i$, $i = 1, 2$. On $(\Omega^*, \mathcal{B}^*, \Pi^*)$ the processes $\Lambda_i^*(f)(\omega^*) := \Lambda_i(f)(U_i(\omega^*))$, $\omega^* \in \Omega^*$, $f \in \Sigma$ then have the same distribution as $\Lambda_i(f)(\omega)$ on $(\Omega_i, \mathcal{B}_i, \Pi_i)$. Define probability measures $P_{i,f}^*$ by $dP_{i,f}^* = \Lambda_i^*(f) d\Pi^*$ and experiments $\mathbb{E}_i^* = (\Omega^*, \mathcal{B}^*, (P_{i,f}^*, f \in \Sigma))$. Then, applying the factorization criterion to the Π^*- densities $\Lambda_i^*(f)$, we see that the maps U_i are sufficient statistics for $\mathbb{E}_i^*$, so that $\mathbb{E}_i^*$ and $\mathbb{E}_i$ are equivalent. This implies the estimate

$$\Delta(\mathbb{E}_1, \mathbb{E}_2) \leq \Delta_0(\mathbb{E}_1^*, \mathbb{E}_2^*).$$

If everything depends on ϵ, for asymptotic equivalence of $\mathbb{E}_{i,\epsilon}$ it then then suffices to prove that $\mathbb{E}_{i,\epsilon}^*$ are asymptotically total variation equivalent, i. e. that

$$\sup_{f \in \Sigma} E_{\Pi^*} \left| \Lambda_{1,\epsilon}^*(f) - \Lambda_{2,\epsilon}^*(f) \right| \to 0, \ \epsilon \to 0. \tag{6}$$

Let $H(\cdot, \cdot)$ be the Hellinger distance between probability measures. A sufficient condition for (6) is

$$\sup_{f \in \Sigma} H^2(P_{1,\epsilon,f}^*, P_{2,\epsilon,f}^*) = \sup_{f \in \Sigma} E_{\Pi^*} \left| (\Lambda_{1,\epsilon}^*(f))^{1/2} - (\Lambda_{2,\epsilon}^*(f))^{1/2} \right|^2 \to 0, \ \epsilon \to 0.$$

This was also the method adopted by Brown and Low (1996), with the Wiener space as natural common probability space for W and ξ_i.
Accordingly, in our diffusion model, consider the density for (1) when the dominating measure is the distribution of ϵW: for $z = \epsilon W$

$$\Lambda_{1,\epsilon}(f) = \exp\left\{ \frac{1}{\epsilon^2} \int_0^1 f(z(t)) dz(t) - \frac{1}{2\epsilon^2} \int_0^1 f^2(z(t)) dt \right\}. \tag{7}$$

For the discrete scheme (5) the analog is for $z_i = \epsilon \, W(t_i)$

$$\Lambda_{2,\epsilon}(f) = \exp\left\{ \frac{1}{\epsilon^2} \sum_{i=1}^n f(z_i)(z_{i+1} - z_i) - \frac{1}{2\epsilon^2 n} \sum_{i=1}^n f^2(z_i) \right\}. \tag{8}$$

To state our main theorem, let us determine the parameter space Σ for our experiments. We assume only the standard condition for existence and uniqueness of a solution y of the SDE (1) (cp. ksendal (1992), theorem 5.5) in a uniform version: let

$$\Sigma_M = \{ f \text{ defined on } R, \ |f(x) - f(u)| \leq M \ |x - u|, \ \ x, u \in R, \ |f(0)| \leq M \}.$$

Observe that $f \in \Sigma_M$ implies the condition of linear growth which is commonly encountered in this connection: $|f(x)| \leq M(1 + |x|)$.

Theorem 1. *Suppose that for some $M > 0$ the parameter space Σ fulfills $\Sigma \subset \Sigma_M$, and that $n = n_\epsilon$ is chosen such that $\epsilon \, n_\epsilon \to \infty$. Then the experiments given by (1) and*

(5) are asymptotically equivalent as $\epsilon \to 0$.

REMARK 1. The model (5) is of the *autoregressive* type, and corresponding diffusion limits have been studied extensively in parametric models. To see the connection, consider the case where the parameter space is

$$\Sigma = \{f,\ f(x) = \vartheta\ x,\ |\vartheta| \le M\} \subset \Sigma_M.$$

Thus we have the parametric model

$$dy(t) = \vartheta y(t)\ dt + \epsilon\ dW(t),\ t \in [0,1] \tag{9}$$

which has been investigated predominantly with fixed ϵ and an increasing interval of observation (see Kutoyants (1984), § 3.5). In our case of fixed interval and varying ϵ all experiments (9) for different ϵ are equivalent (exactly). Indeed, multiplying the observations y by ϵ^{-1} yields an equivalent experiment, and the process $\tilde{y} = \epsilon^{-1}y$ satisfies (9) for $\epsilon = 1$. Thus our accompanying sequence of experiments is constant with respect to type, and theorem 1 for $n = \epsilon^{-2}$ establishes a *strong* $(\Delta-)$ *convergence* of the autoregression experiments

$$y_{i+1} = y_i + n^{-1}\vartheta y_i + n^{-1}\xi_i,\ i = 1,\dots,n,\ y_1 = 0 \tag{10}$$

to a diffusion limit (9) with $\epsilon = 1$. Define $\tilde{y}_i = ny_i$ and $\beta = 1 + n^{-1}\vartheta$; then (10) may be written

$$\tilde{y}_{i+1} = \beta\ \tilde{y}_i + \xi_i,\ i = 1,\dots,n,\ y_1 = 0 \tag{11}$$

which is the familiar AR(1) model in the *nearly nonstationary* case where the parameter β is close to 1. Parametric inference in these models based on the diffusion limit has been studied by Chan and Wei (1987), Cox (1991); cp. also comments in Jeganathan (1988), and Stockmarr and Jacobsen (1994) for a multivariate AR(1) model. We have thus recovered the limit experiment argument for critical autoregression; the behaviour is known as "locally asymptotically Brownian functional" (Phillips, 1987)) or locally asymptotically quadratic (Le Cam and Yang, 1990). For a comprehensive discussion cf. Shiryaev and Spokoiny (1996), chap V. Thus theorem 1 appears as a natural nonparametric extension of the LAQ limit, the limit being substituted by accompanying experiments.

REMARK 2. It is now clear what the relation to *nonparametric autoregression* should be: define $\tilde{y}_i$ as above, $n = \epsilon^{-2}$ and a function $g(x) = x + f(n^{-1}x)$; then (5) may be written

$$\tilde{y}_{i+1} = g(\tilde{y}_i) + \xi_i,\ i = 1,\dots,n,\ y_1 = 0. \tag{12}$$

Nonparametric inference for fixed, unknown g was studied by Doukhan and Ghindès (1983) under stationarity assumptions. They found that the theory of kernel type estimators parallels the signal plus white noise case (2), as regards rates of convergence. So obviously the nonparametric model (12) with g fixed and stationarity corresponds to parametric autoregression in the stable case where β is bounded away from 1 (and local asymptotic normality holds), while in the nearly critical case $g(x) = x + f(n^{-1}x)$, $f \in \Sigma$ the diffusion approximation of theorem 1 holds.

REMARK 3. Parametric asymptotic results for autoregressive models are naturally available for nonnormal ξ_i, based on the limit experiment rationale. This raises the question whether theorem 1 might be valid also for nongaussian ξ_i in the autoregression. Recent results for the regression model (3) under nongaussian errors (Grama and Nussbaum (1996)) do suggest a positive answer.

Theorem 1 is proved with a reasoning related to the one of Brown and Low (1996) in the 'signal plus noise' case. The two models (1) and (5), even when they are construed as being on the same probability space, are still of different kind: one is discrete, the other continuous. To facilitate the treatment, it is convenient to compare the likelihoods of two continuous models. That is achieved by *interpolation* of the discrete process y_i, $i = 1, \ldots, n$. Recall the classical Euler scheme for solving the (deterministic) differential equation

$$\frac{d}{dt} y^0(t) = f(y^0(t)),\ t \in [0,1] \tag{13}$$

which is (5) without the random term, i. e.

$$n(y^0_{i+1} - y^0_i) = f(y^0_i),\ i = 1, \ldots, n,\ y^0_1 = 0. \tag{14}$$

The rationale is of course closely linked to linear interpolation: if $\bar{y}^0$ is a piecewise linear function on $[0,1]$ such that $\bar{y}^0(t_i) = y_i$ then its derivative on $[t_i, t_{i+1}]$ is $f(\bar{y}^0(t_i))$. The classical Euler scheme indeed consists in taking the linear interpolation $\bar{y}^0$ of the values y^0_i as the approximate solution of the ordinary differential equation.
For the stochastic Euler scheme (5) one might also consider a linear interpolation of the values y_i. But in order to account for the oscillating behaviour of the solution y of (1) inherited from the driving Wiener process, one might then *add a Brownian bridge* to the linear interpolation on each interval $[t_i, t_{i+1}]$. These Brownian bridges should be independent of the ξ_i; call this randomly interpolated process $\bar{y}$.
A convenient representation of $\bar{y}$ can be obtained as follows. Define a function $\bar{f}_n$ on $[0,1]$ which depends on a path $z(t)$, $t \in [0,1]$ as

$$\bar{f}_n(t, z) = \sum_{i=1}^{n} f(z(t_i))\ \chi_{(t_i, t_{i+1}]}(t).$$

Thus $\bar{f}_n$ is a piecewise constant function which interpolates $f(z(\cdot))$ in t_i.

Lemma 1. *The unique solution $\bar{y}$ of*

$$\bar{y}(t) = \int_0^t \bar{f}_n(u, \bar{y})\, du + \epsilon\, W(t) \tag{15}$$

may be represented as a randomly interpolated process as above, for an appropriate choice of the n Brownian bridges.

The proof is in section 4. It is clear that $\bar{y}$ contains as much information about f as $y_i,\ i = 1, \ldots, n$. Indeed $\bar{y}(t_i) = y_i$, but on the other hand linear interpolation and adding Brownian bridges which do not depend on f do not increase the informational content. Thus the experiments given by (5) and (15) are equivalent, for any parameter space Σ, and $\bar{y}(t_i),\ i = 1, \ldots, n$ is a sufficient statistic in (15). This can also be seen by looking at the likelihood for model (15) as follows.

Regard $\bar{y}$ formally as a diffusion type process defined by (15), see Liptser, Shiryaev, (1974), Chap. 4, §2, Definition 7. Indeed $\bar{f}_n(u, \bar{y})$ is for each u a non-anticipating functional, since it depends on $\bar{y}$ only via $\bar{y}(t_i)$, $u \in (t_i, t_{i+1}]$. (But $\bar{y}$ is not a diffusion process in the strict sense, which would mean that $\bar{f}_n(u, \bar{y})$ depends $\bar{y}$ only via $\bar{y}(u)$). This process has a distribution which is absolutely continuous with respect to the distribution of $\epsilon\ W$ if almost surely $\int_0^1 \bar{f}_n^2(u, \bar{y}) du < \infty$ (Liptser, Shiryaev, (1974), Chap. 7, §2, Theorem 7.6). But this is fulfilled since every value y_i from (5) is finite. Then the density is analogously to (7) for $z = \epsilon W$

$$\Lambda_{3,\epsilon}(f) = \exp\left\{ \frac{1}{\epsilon^2} \int_0^1 \bar{f}_n(t, z)\, dz(t) - \frac{1}{2\epsilon^2} \int_0^1 \bar{f}_n^2(t, z)\, dt \right\}. \tag{16}$$

Now observe

$$\int_0^1 \bar{f}_n(t, z)\, dz(t) = \sum_{i=1}^n f(z_i)(z_{i+1} - z_i),\ \int_0^1 \bar{f}_n^2(t, z)\, dt = \sum_{i=1}^n f^2(z_i) n^{-1}$$

so that we obtain $\Lambda_{3,\epsilon}(f) = \Lambda_{2,\epsilon}(f)$. This means that the density (16) depends on z via $z_i = z(t_i)$ only, which again implies that the values $\bar{y}(t_i),\ i = 1, \ldots, n$ are a sufficient statistic in (15) and the experiments given by (5) and (15) are equivalent, for any parameter space Σ.

It remains to establish that the densities of the two diffusion type processes (1) and (15) fulfill

$$\sup_{f \in \Sigma} E \left| \Lambda_{1,\epsilon}^{1/2}(f) - \Lambda_{3,\epsilon}^{1/2}(f) \right|^2 \to 0,\ \epsilon \to 0. \tag{17}$$

For the signal plus noise case (2) Brown and Low (1996) were able to use the explicit formula for the squared Hellinger distance $H^2(\cdot, \cdot)$ of two shifted Wiener measures: if $Q_{\epsilon,f}$, Q_{ϵ,f^*} are two measures corresponding to (2) with drift functions f, f^*, say, then

$$H^2(Q_{\epsilon,f}, Q_{\epsilon,f^*}) = 2 - 2 \exp\left\{ -\frac{1}{8} \epsilon^{-2} \left\| f - f^* \right\|^2 \right\} \tag{18}$$

(in analogy to the finite dimensional Gaussian case). In particular f^* was taken as a step function approximant for f. In our more involved diffusion model, we use an inequality derived from the *Hellinger process* in semimartingale theory, cf. Jacod and Shiryaev (1987). Denote by $P_{\epsilon,f}$, $P^*_{\epsilon,f}$ the probability measures on $C[0,1]$ given by the two processes (1) and (15), respectively; the notation $P^*_{\epsilon,f}$ reflects the dependence of $\bar{f}_n(\cdot,\cdot)$ in (15) on f. Let h_f be the Hellinger process of order 1/2 between these two measures (see Jacod and Shiryaev (1987), §4b, p. 239): for a realization z

$$h_f(u) = \frac{1}{8\epsilon^2}\int_0^u (f(z(t)) - \bar{f}_n(t,z))^2 dt.$$

The inequality for the the total variation distance $\|\cdot\|_{TV}$ is

$$\left\|P_{\epsilon,f} - P^*_{\epsilon,f}\right\|_{TV} (= E_{\epsilon,0}\,|\Lambda_{1,\epsilon}(f) - \Lambda_{3,\epsilon}(f)|) \le 4\sqrt{E_{\epsilon,f}\; h_f(1)}, \tag{19}$$

where $E_{\epsilon,f}$ denotes expectation wrt $P_{\epsilon,f}$, see Jacod and Shiryaev (1987), §4b, theorem 4.21, p. 279. A remarkable feature of inequality (19) is that it is nonsymmetric and holds both ways with expectation taken with respect to either $P_{\epsilon,f}$ or $P^*_{\epsilon,f}$. The choice $P_{\epsilon,f}$ somewhat facilitates evaluation of $E_{\epsilon,f}\; h_f$ which is our next task.

Lemma 2. *Suppose Σ is as in theorem 1. Then*

$$E_{\epsilon,f}\int_0^1 (f(y(t)) - \bar{f}_n(t,y))^2 dt = O(n^{-2} + \epsilon^2 n^{-1}),\ \epsilon \to 0$$

uniformly over $f \in \Sigma$, where $E_{\epsilon,f}$ denotes expectation with respect to the distribution $P_{\epsilon,f}$ of the process y in (1).

The proof is in section 4. We obtain

$$\left\|P_{\epsilon,f} - P^*_{\epsilon,f}\right\|_{TV} = O(\epsilon^{-1}n^{-1} + n^{-1/2})$$

uniformly, which establishes theorem 1.

3 Sampling from a Diffusion

Let us now consider the situation where one has discrete data from the diffusion process y in (1). Various questions of inference based on sampled values $y(t_i), i = 1, \ldots, n$ have been treated in the literature: Prakasa Rao (1983), Dacunha-Castelle and Florens-Zmirou (1986), Florens-Zmirou (1989), Yoshida (1992), Kloeden, Schurz, Platen and Srensen (1993).

Theorem 2. *Under the conditions of theorem 1, in the diffusion model (1) with $\epsilon \to 0$ the sampled values $y(t_1), \ldots, y(t_n)$ are an asymptotically sufficient statistic.*

Proof. Let $\mathbb{E}_{1,\epsilon}$ be the experiment given by observations y in (1) and $\mathbb{E}_{3,\epsilon}$ be given by $\bar{y}$ in (15). Define for a path $z(t),\ t \in [0,1]$ the statistic T_ϵ by $T_\epsilon(z) = (z(t_1), \ldots, z(t_n))$. Then T_ϵ is sufficient in $\mathbb{E}_{3,\epsilon}$, and the result follows from (17). $\Box$

Laredo (1990) obtained asymptotic sufficiency of the sampled values in a parametric framework using a different method, based on the exponential family approximation related to the standard local asymptotic normality results. The asymptotic sufficiency then naturally is a local one, i. e. it refers to experiments restricted to neighborhoods shrinking at rate ϵ. Though the result may be globalized using preliminary estimators, this methodology fails in the nonparametric case since rate ϵ estimators are not available. In contrast our result in the nonparametric case is global over $f \in \Sigma$ from the outset. In the parametric local approach other statistics like level crossings have also been treated, see Genon-Catalot and Larédo (1990).

4 Proofs of lemmas

Proof of lemma 1. Recall that $\xi_i = n^{1/2}(W(t_{i+1}) - W(t_i))$; define

$$B_i(t) = W(t) - W(t_i) - n(t - t_i)\ (W(t_{i+1}) - W(t_i)) \text{ for } t \in [t_i, t_{i+1}]$$

and $B_i(t) = 0$ otherwise; thus $B_i(t)$ is a Brownian bridge on $[t_i, t_{i+1}]$ which is independent of ξ_i. Hence the $B_i(t),\ i = 1, \ldots, n$ are mutually independent and independent of the $\xi_i,\ i = 1, \ldots, n$. Now we have $\bar{y}(0) = 0 = y_1$ by definition of the interpolated process and also by (15). Thus $\bar{y}(t_1) = y_1$, and for the induction step we may assume $\bar{y}(t_i) = y_i$. Then on $[t_i, t_{i+1}]$ we get

$$\bar{y}(t) = y_i + \int_{t_i}^{t} \bar{f}_n(u, \bar{y})\ du + \epsilon\ (W(t) - W(t_i)) = y_i + (t - t_i)(f(y_i) + \epsilon\ n^{1/2}\xi_i) + \epsilon\ B_i(t).$$

Hence $\bar{y}(t_{i+1}) = y_i + n^{-1}f(y_i) + \epsilon\ n^{-1/2}\xi_i = y_{i+1}$ and on $[t_i, t_{i+1}]$ the solution $\bar{y}(t)$ is

$$\bar{y}(t) = y_i + n(t - t_i)y_{i+1} + \epsilon\ B_i(t),$$

i. e. it is the linear interpolation of y_i and y_{i+1} plus a Brownian bridge.

Proof of lemma 2. We have

$$\int_0^1 (f(y(t)) - \bar{f}_n(t, y))^2 dt = \sum_{i=1}^{n} \int_{t_i}^{t_{i+1}} (f(y(t)) - f(y(t_i))^2\ dt$$

$$\leq \sum_{i=1}^{n} n^{-1} \sup_{t \in (t_i, t_{i+1}]} M\ |(y(t) - y(t_i)|^2$$

$$\leq 2n^{-1} \sum_{i=1}^{n} \sup_{t \in (t_i, t_{i+1}]} \left| \int_{t_i}^{t} f(y(u))\ du \right|^2 + 2n^{-1} \sum_{i=1}^{n} \sup_{t \in (t_i, t_{i+1}]} \epsilon^2\ |W(t) - W(t_i)|^2 . \tag{20}$$

For the first term in (20), apply Cauchy-Schwartz and $|f(x)| \le M(1+|x|)$ to obtain

$$n^{-1} \sum_{i=1}^{n} \sup_{t \in (t_i, t_{i+1}]} \left| \int_{t_i}^{t} f(y(u))\, du \right|^2 \le n^{-1} \sum_{i=1}^{n} n^{-1} \int_{t_i}^{t_{i+1}} f^2(y(u))\, du$$

$$\le 2n^{-2} \sum_{i=1}^{n} \int_{t_i}^{t_{i+1}} M^2(1+y^2(u))du. \tag{21}$$

To estimate the expectation of this term, we use the following relation: there is a constant C_M depending on M but not on ϵ such that for all $f \in \Sigma$

$$E_f\ y^2(t) \le C_M,\ \ t \in [0,1].$$

Indeed this follows from ksendal (1992), exercise 5.6, which is connected with the existence and uniqueness theorem for SDE (note that in (1) we have an SDE with diffusion coefficient $\sigma(t,y) = \epsilon$). Taking an expectation in (21) we get

$$E_f n^{-2} \sum_{i=1}^{n} \int_{t_i}^{t_{i+1}} M^2(1+y^2(u))du \le n^{-2} M^2(1+C_M). \tag{22}$$

We have estimated the expectation of the first term in (20) as $O(n^{-2})$ uniformly. For the second term, note that this is an average of identically distributed random variables, which have the distribution of

$$\sup_{t \in [0,n^{-1})} \epsilon^2 |W(t)|^2 \simeq \epsilon^2 n^{-1} \sup_{t \in [0,1)} |W(t)|^2$$

where '$\simeq$' means equality in law. But $\sup_{t \in [0,1)} |W(t)|^2$ has finite expectation (Breiman (1968), ch. 13.7), so that the expectation of the second term in (20) is $O(\epsilon^2 n^{-1})$ uniformly.

ACKNOWLEDGEMENT. The authors wish to thank Vladimir Spokoiny and Priscilla Greenwood for helpful discussions.

References

[1] Breiman, L. (1968). *Probability.* Addison-Wesley, New York.

[2] Brown, L. D. and Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise. To appear, Ann. Statist.

[3] Chan, N. H. and Wei, C. Z. (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Ann. Statist.* **15** 1050-1063.

[4] Cox, D. D. (1991). Gaussian likelihood estimation for nearly nonstationary AR(1) processes. *Ann. Statist.* **19** 1129-1142

[5] Dacunha-Castelle, D. and Florens-Zmirou, D. (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics* **19** 263-284

[6] Doukhan, P. and Ghindès, M. (1983). Estimation de la probabilité de transition d'une chaîne de Markov Doëblin-récurrente. Etude du cas du processus autorégressif général d'ordre 1. *Stochastic Processes Appl.* **15** 271-292

[7] Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics* **20** (4) 547-557

[8] Genon-Catalot, V., Larédo, C. (1990). An asymptotic sufficiency property of observations related to the first hitting times of a diffusion. *Statistics & Probab. Letters* . **9** 253-258

[9] Grama, I. and Nussbaum, M. (1996). Asymptotic equivalence of regression experiments. Forthcoming.

[10] Höpfner, R. , Jacod, J., Ladelli, L. (1990). Local asymptotic normality and mixed normality for Markov statistical models. Probab. Th. Rel. Fields 86 105-129

[11] Jacod, J., Shiryaev, A. (1987). *Limit theorems for Stochastic Processes*. Springer-Verlag, Berlin, Heidelberg, New York.

[12] 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** 1283-1314

[13] Kloeden, P. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations.* Springer-Verlag, Berlin, Heidelberg, New York.

[14] Kloeden, P., Schurz, H., Platen, E. and Srensen, M. (1993). On effects of discretization of estimators of drift parameters for diffusion processes. Submitted.

[15] Korostelev, A. P. and Nussbaum, M. (1996). The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. Discussion paper, SFB 373, Humboldt University Berlin, submitted.

[16] Kutoyants, Yu. A. (1984). *Parameter Estimation for Stochastic Processes.* Heldermann, Berlin

[17] Kutoyants, Yu. A. (1985). On nonparametric estimation of trend coefficients in a diffusion process. In: *Statistics and Control of Stochastic Processes, Steklov Seminar 1984* (Krylov, _N. V., Liptser, R. Sh., Novikov, A. A. Eds.), Optimization Software, Inc, New York

[18] Laredo, C. F. (1990). A sufficient condition for asymptotic sufficiency of incomplete observations of a diffusion process. *Ann. Statist.* **18** 1158-1171

[19] Liptser, R. S., Shiryaev, A. (1977). *Statistics of Stochastic Processes.* Springer-Verlag, Berlin, Heidelberg, New York.

[20] Le Cam, L. (1985). Sur l'approximation de familles de mesures par des familles gaussiennes. *Ann. Inst. Henri Poincaré* **21** (3) 225-287

[21] Le Cam , L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, Berlin, Heidelberg, New York.

[22] Le Cam , L. and Yang, G. (1990). *Asymptotics in Statistics.* Springer-Verlag, Berlin, Heidelberg, New York.

[23] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. To appear, Ann. Statist.

[24] Phillips, P. C. B. (1987). Towards a unified asymptotic theory for autoregression. Biometrika **74** 535-547

[25] ksendal, B. (1992). *Stochastic Differential Equations.* Springer-Verlag, Berlin, Heidelberg, New York.

[26] Prakasa Rao, B. L. S. (1983). Statistical inference from sampled data for stochastic processes. *Contemporary Mathematics*, Vol. 80, 249-284. Amer. Math. Soc., Providence.

[27] Shiryaev, A. N., Spokoiny, V. (1993). *Statistical Experiments and Decisions: Asymptotic Theory.* To appear.

[28] Stockmarr, A. and Jacobsen, M. (1994). Gaussian diffusions and autoregressive processes: weak convergence and statistical inference. Scand. J. Statist. **21** 4 403-419

[29] Yoshida, N. (1992). Estimation for diffusion processes from discrete observations. J. Multivariate Analysis **41**, 220-242

WEIERSTRASS INSTITUTE
MOHRENSTR. 39
D-10117 BERLIN
GERMANY
E-MAIL MILSTEIN@WIAS-BERLIN.DE;
NUSSBAUM@WIAS-BERLIN.DE