Efficient estimation for ergodic diffusion processes sampled at high frequency

Michael Sørensen
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5, DK-2100 Copenhagen Ø, Denmark

January 9, 2024

Abstract

A general theory of efficient estimation for ergodic diffusion processes sampled at high frequency with an infinite time horizon is presented. High frequency sampling is common in many applications, with finance as a prominent example. The theory is formulated in term of approximate martingale estimating functions and covers a large class of estimators including most of the previously proposed estimators for diffusion processes. Easily checked conditions ensuring that an estimating function is an approximate martingale are derived, and general conditions ensuring consistency and asymptotic normality of estimators are given. Most importantly, simple conditions are given that ensure rate optimality and efficiency. Rate optimal estimators of parameters in the diffusion coefficient converge faster than estimators of drift coefficient parameters because they take advantage of the information in the quadratic variation. The conditions facilitate the choice among the multitude of estimators that have been proposed for diffusion models. Optimal martingale estimating functions in the sense of Godambe and Heyde and their high frequency approximations are, under weak conditions, shown to satisfy the conditions for rate optimality and efficiency. This provides a natural feasible method of constructing explicit rate optimal and efficient estimating functions by solving a linear equation.

Key words: Approximate martingale estimating functions, discrete time observation of a diffusion process, efficiency, Euler approximation, explicit estimating functions, generalized method of moments, optimal estimating function, optimal rate, maximum likelihood estimation, stochastic differential equations.

1 Introduction

Dynamic phenomena affected by random noise are often modelled in continuous time by stochastic differential equations. Among the advantages of this approach are interpretable model parameters and easy communication with other scientists by using a common modelling tool, viz. differential equations. A few examples are applications in the areas of animal movement (Michelot *et al.* (2019)), climate research (Ditlevsen and Ditlevsen (2023)), finance (Chan *et al.* (1992), Dipple *et al.* (2020)), protein structure evolution (Golden *et al.* (2017)), neuro science (Bibbona *et al.* (2010), Jensen *et al.* (2012), Bachar *et al.* (2013)), transmission of infectious diseases (Guy *et al.* (2015), Arnst *et al.* (2022)) and physiology (Picchini *et al.* (2008)). While the dynamics is formulated in continuous time, observations are made at discrete points in time. This complicates statistical inference for these models, which is an intensive area of research, where a profusion of estimators have been proposed for parametric diffusion models, see e.g. Sørensen (2012) and Iacus and Yoshida (2018). Many simulation studies have been performed to compare the relative merits of estimators, but have not provided a clear general picture. The simple and easily checked criteria for efficiency and rate optimality of estimators obtained in this paper are useful for identifying the best estimators and explain findings in simulation studies.

We consider a scalar diffusion given by the stochastic differential equation

$$dX_t = b(X_t; \alpha)dt + \sigma(X_t; \beta)dW_t, \tag{1.1}$$

where $(\alpha, \beta) = \theta \in \Theta \subseteq \mathbb{R}^2$ are parameters to be estimated. The restrictions to a scalar process and to two scalar parameters are made to simplify the presentation. The results can be generalized to multivariate diffusions as indicated in Section 4, and to multivariate parameters by considering estimating functions of the same dimension as the parameter vector and replacing partial derivatives by vectors or matrices of partial derivatives. The process X is assumed to be observed at equidistant time points $i\Delta_n$, $i = 0, \ldots, n$, and we consider the high frequency/infinite time horizon asymptotic scenario, where

$$n \to \infty, \qquad \Delta_n \to 0, \qquad n\Delta_n \to \infty.$$

The length of the time interval in which observations are made goes to infinity, which is necessary to ensure that the drift parameter α can be estimated consistently. At the same time the sampling frequency goes to infinity. This is particularly important for diffusion processes, because the quadratic variation of a diffusion process contains information about the parameter β in the diffusion coefficient, which good estimators can use when the sampling frequency is sufficiently high. For such estimators, β is estimated at a higher rate than the $1/\sqrt{n\Delta_n}$-rate, which is optimal for the drift parameter α; see Gobet (2002) where it is shown that the model (1.1) is locally asymptotically normal under the high frequency/infinite time horizon asymptotic scenario. In the present paper we give easily checked conditions for rate optimality and efficiency. If the drift coefficient is known, consistent estimators can be found without the infinite time horizon assumption. Results on rate optimality and efficiency when the sampling interval is bounded are given in Jakobsen and Sørensen (2017). High frequency asymptotics is often relevant in applications, because the sampling frequency needs not be particularly high for the asymptotic optimality results to work. It only needs to be high relative to the characteristic time scale of the diffusion process. For some types of economic data, even weekly observations can in be considered a high sampling frequency, see e.g. Larsen and Sørensen (2007).

Our theory is phrased in terms of estimating functions of the general form

$$G_n(\theta) = \sum_{i=1}^{n} g(\Delta_n, X_{i\Delta_n}, X_{(i-1)\Delta_n}; \theta), \tag{1.2}$$

where the function $g(\Delta, y, x; \theta)$, with values in $\mathbb{R}^2$, is such that G_n is, exactly or approximately, a martingale estimating function. Specifically, $E_\theta(g(\Delta, X_{i\Delta_n}, X_{(i-1)\Delta_n}; \theta) \mid X_{(i-1)\Delta_n})$ is equal to zero or of order Δ^κ for some $\kappa \geq 2$. Estimators are obtained as solutions to the estimating equation $G_n(\theta) = 0$. We call such an estimator a *G_n-estimator*. For estimating functions that are not exact martingales, the extra condition that $n\Delta^{2(\kappa-1)} \to 0$ is needed to ensure the asymptotic results.

The theory developed here covers a large class of estimators for diffusion processes including most of the previously proposed estimators. The few that are not covered are likely to be less efficient, because non-martingale estimating functions, in general, do not approximate the score function as well as martingales. In particular, the theory covers the martingale estimating functions proposed by Bibby and Sørensen (1995) and Kessler and Sørensen (1999), GMM-estimators based on conditional moments, Hansen (1982, 1985, 1993), and the maximum likelihood estimator and Bayesian estimators; for numerical methods to calculate these estimators, see Pedersen (1995), Roberts and Stramer (2001), Aït-Sahalia (2002), Durham and Gallant (2002), Aït-Sahalia and Mykland (2003), Beskos *et al.* (2009), Golightly and Wilkinson (2011), Bladt and Sørensen (2014), van der Meulen and Schauer (2017) and Bladt *et al.* (2021). The pseudo-likelihood function obtained from the Gaussian Euler approximation to the transition density is covered too. Estimators closely related to the Euler pseudo-likelihood were considered by Florens-Zmirou (1989), Yoshida (1992) and Uchida (2010). These and pseudo-likelihood functions based on more accurate Gaussian approximations to the likelihood function, such as those considered by Kessler (1997), Uchida and Yoshida (2012) and Kitagawa and Uchida (2014), are also covered. Sørensen and Uchida (2003) and Gloter and Sørensen (2009) considered the Euler pseudo likelihood under a combination of high frequency and small diffusion asymptotics, where the diffusion coefficient goes to zero as $n \to \infty$. The latter condition replaces the infinite time horizon condition.

The following condition on the function $g(\Delta, y, x; \theta)$ ensures *rate optimality* of estimators.

Condition 1.1

$$\partial_y g_2(0, x, x; \theta) = 0 \tag{1.3}$$

for all x in the state-space of the diffusion process and all $\theta \in \Theta$.

By $\partial_y g_2(0, x, x; \theta)$ we mean $\partial_y g_2(0, y, x; \theta)$ evaluated at $y = x$. Here ∂_y denotes the partial derivative w.r.t. y, and g_i is the ith coordinate of g. More precisely, it is sufficient that a linear combination of the two coordinates of g satisfies (1.3), but without loss of generality it can be assumed to be g_2. This will be explained in Section 2. We will refer to (1.3) as Jacobsen's condition because it equals one of the conditions for small Δ-optimality in the sense of Jacobsen (2001) of martingale estimating functions; see Jacobsen (2002). Jacobsen considered an asymptotic scenario where the time between observations Δ does not depend on n. In his approach the Condition 1.1 was introduced to avoid a singularity in the asymptotic variance of the estimators at $\Delta = 0$. In our high frequency approach, the condition implies rate optimality for estimation of the diffusion coefficient parameter.

Our condition for *efficiency* is

Condition 1.2

$$\partial_y g_1(0, x, x; \theta) = \partial_\alpha b(x; \alpha)/\sigma^2(x; \beta) \tag{1.4}$$

and

$$\partial_y^2 g_2(0, x, x; \theta) = \partial_\beta \sigma^2(x; \beta)/\sigma^4(x; \beta), \tag{1.5}$$

for all x in the state space of the diffusion process and all $\theta \in \Theta$.

The Conditions 1.1 and 1.2 are, under weak regularity conditions, shown to be satisfied by martingale estimating functions that are optimal in the sense of Godambe and Heyde (1987), which is both very useful and quite surprising. Useful because it provides an easy method of constructing rate optimal and efficient estimators, and surprising because Godambe-Heyde optimality is a local property in the sense that it is a property of a particular class of estimating functions. Therefore, there is no a priori reason to except this property to imply global properties like rate optimality and efficiency. Martingale estimating functions give consistent estimators at all sampling frequencies, see Bibby *et al.* (2010), and Godambe-Heyde optimal martingale estimating functions are known to often provide estimators with a high efficiency, see e.g. the simulation studies in Overbeck and Rydén (1997) and Larsen and Sørensen (2007). The results in this paper explain why this is the case.

The paper is organized as follows. Section 2 sets up the model, the class of approximate martingale estimating functions, and the assumptions and the notation used throughout the paper. A number of well-known estimators are shown to be covered by the theory, and a lemma of independent interest gives fundamental identities and characterizes approximate martingale estimating functions of order κ. Section 3 develops the high frequency asymptotic theory for general estimating functions as well as for estimating functions satisfying Condition 1.1. The conditions for efficiency are derived in Section 4, and it is proved that Godambe-Heyde optimal martingale estimating functions and their high frequency approximations are rate optimal and efficient, which provides a feasible method of constructing explicit rate optimal and efficient estimating functions. Examples are considered, including the Euler pseudo-likelihood and maximum likelihood estimation. Proofs and some lemmas are given in Section 5, where also tools for studying high frequency asymptotic properties of estimators are provided. Section 6 concludes.

2 Model, conditions and notation

We consider observations $X_{t_0^n}, \ldots, X_{t_n^n}$ of the process given by (1.1) at the time points $t_i^n = i\Delta_n$, $i = 0, \ldots, n$. We suppose that a solution of the stochastic differential equation (1.1) exists, is unique in law, and is adapted to the filtration generated by the Wiener process W and the initial value X_0. The state-space of X is denoted by (ℓ, r), where $-\infty \leq \ell < r \leq \infty$, and we assume that $v(x; \beta) = \sigma^2(x; \beta) > 0$ for all $x \in (\ell, r)$. Furthermore, we assume that $\theta = (\alpha, \beta) \in \Theta$, where Θ is a subset of $\mathbb{R}^2$, and that the true parameter value $\theta_0 = (\alpha_0, \beta_0) \in \operatorname{int}\Theta$, the interior of Θ. It is no serious restriction to assume that Θ is convex.

A function $f(y, x; \theta)$ is said to be of *polynomial growth* in y and x uniformly for θ in a compact set if, for any compact subset $K \subseteq \Theta$, there exists a constant $C > 0$ such that $\sup_{\theta \in K} |f(y, x; \theta)| \leq C(1 + |x|^C + |y|^C)$ for all $x, y \in (\ell, r)$. The assumptions of polynomial growth in this paper are made to simplify the theory. These assumptions are satisfied for most models used in practice, but could be relaxed.

Here and in the rest of the paper, $R(\Delta, y, x; \theta)$ denotes a (generic) function such that $|R(\Delta, y, x; \theta)| \le F(y, x; \theta)$, for all Δ, where F is some function of polynomial growth in y and x uniformly for θ in a compact set. Similarly for $R(\Delta, x; \theta)$.

Definition 2.1 *We let $C_{p,k_1,k_2,k_3}(\mathbb{R}_+ \times (\ell, r)^2 \times \Theta)$ denote the class of real functions $f(t, y, x; \theta)$ satisfying that*

(i) $f(t, y, x; \theta)$ is k_1 times continuously differentiable with respect to t, k_2 times continuously differentiable with respect to y, and k_3 times continuously differentiable with respect to θ

(ii) f and all partial derivatives $\partial_t^{i_1} \partial_y^{i_2} \partial_\alpha^{i_3} \partial_\beta^{i_4} f$, $i_j = 0, \ldots k_j$, $j = 1, 2$, $i_3 + i_4 \le k_3$ are continuously differentiable with respect to x and are of polynomial growth in x and y uniformly for θ in compact sets (for fixed $t \le 1$)

(iii) f has an expansion

$$f(\Delta, y, x; \theta) = \sum_{i=0}^{k_1} \frac{\Delta^i}{i!} f^{(i)}(y, x; \theta) + \Delta^{k_1+1} R(\Delta, y, x; \theta). \tag{2.1}$$

The classes $C_{p,k_2,k_3}((\ell, r) \times \Theta)$ and $C_{p,k_2,k_3}((\ell, r)^2 \times \Theta)$ are defined similarly (with property (iii) omitted) for functions of the form $f(y; \theta)$ and $f(y, x; \theta)$, respectively.

We assume that the stochastic differential equation (1.1) satisfies the following condition.

Condition 2.2 *The following holds for all $\theta \in \Theta$:*

(1)

$$\int_{x^\#}^{r} s(x; \theta) dx = \int_{\ell}^{x^\#} s(x; \theta) dx = \infty \tag{2.2}$$

and

$$\int_{\ell}^{r} x^k \tilde{\mu}_\theta(x) dx < \infty \tag{2.3}$$

for all $k \in \mathbb{N}$, where $x^\#$ is an arbitrary point in (ℓ, r),

$$s(x; \theta) = \exp\left(-2 \int_{x^\#}^{x} \frac{b(y; \alpha)}{v(y; \beta)} dy \right) \tag{2.4}$$

and

$$\tilde{\mu}_\theta(x) = [s(x; \theta) v(x; \beta)]^{-1} \tag{2.5}$$

(2) $\sup_t E_\theta(|X_t|^k) < \infty$ for all $k \in \mathbb{N}$

(3) $b, \sigma \in C_{p,4,1}((\ell, r) \times \Theta)$

(4) There exists a constant C_θ such that for all $x, y \in (\ell, r)$

$$|b(x;\alpha) - b(y;\alpha)| + |\sigma(x;\beta) - \sigma(y;\beta)| \leq C_\theta |x - y|.$$

The conditions (2.2) and (2.3) with $k = 1$ ensure that the process X is ergodic with an invariant probability measure with Lebesgue density

$$\mu_\theta(x) = \tilde{\mu}_\theta(x) / \int_\ell^r \tilde{\mu}_\theta(y) dy. \tag{2.6}$$

If X is stationary, Condition 2.2 (2) is obviously satisfied under (2.3). Similarly, if X is sufficiently mixing that all moments converge as $t \to \infty$.

We consider estimating functions of the general form (1.2) where the function $g(\Delta, y, x; \theta)$ has values in $\mathbb{R}^2$ and satisfies the following condition.

Condition 2.3

(1) There exists a $\kappa \geq 2$ such that

$$E_\theta(g_i(\Delta_n, X_{t_j^n}, X_{t_{j-1}^n}; \theta) \,|\, X_{t_{j-1}^n}) = \Delta_n^\kappa R(\Delta_n, X_{t_{j-1}^n}; \theta) \quad \textit{for } i = 1, 2 \textit{ and all } \theta \in \Theta \tag{2.7}$$

(2)

$$\begin{aligned} g_i(\Delta, y, x; \theta) &\in C_{p,2,6,2}(\mathbb{R}_+ \times (\ell, r)^2 \times \Theta), \; i = 1, 2 \\ g_i^{(j)}(y, x; \theta) &\in C_{p,2(3-i),2}((\ell, r)^2 \times \Theta), \; i = 1, 2, \; j = 0, 1, 2, \end{aligned}$$

where the $g^{(j)}$s are the functions appearing in the expansion (2.1).

We call an estimating function satisfying Condition 2.3 (1) an *approximate martingale estimating function of order κ*. For any non-singular 2×2 matrix, $M(\Delta, n, \theta)$, the estimating functions $M(\Delta_n, n, \theta) G_n(\theta)$ and $G_n(\theta)$ give identical estimators. We call them *versions* of the same estimating function. Since the matrix M may depend on Δ_n, not all versions satisfy Condition 2.3 and other conditions in the paper, in particular Conditions 1.1 and 1.2. We say that (1.2) is an approximate martingale estimating function of order κ, if there exists a version which satisfies (2.7), and for which the limit $g(0, y, x; \theta)$ is finite with, for at least one value of (x, y, θ), all coordinates different from zero. We use this version in the proofs. It is, typically, obtained by multiplying one or both of the coordinates by a power of Δ_n; examples are given in Section 4.

It could be argued, that an estimating functions that satisfies (2.7) with $\kappa = 1$ could equally well be called an approximate martingale. However, the asymptotic theory in this case is entirely different from the case $\kappa \geq 2$ and requires a separate study. Some particular examples are studied in Jørgensen and Sørensen (2021).

The generator of the solution to (1.1) is the differential operator

$$L_\theta = b(x;\alpha) \frac{d}{dx} + \frac{1}{2} v(x;\beta) \frac{d^2}{dx^2} \tag{2.8}$$

Here we take the domain of L_θ to be the set of all twice continuously differentiable functions defined on the state space. For $f \in C_{p,2(k+1)}((\ell, r))$ and $b, \sigma \in C_{p,2k,0}((\ell, r) \times \Theta)$,

$$\mathrm{E}_\theta(f(X_{t+\Delta}) \,|\, X_t) = \sum_{i=0}^k \frac{\Delta^i}{i!} L_\theta^i f(X_t) + \Delta^{k+1} R(\Delta, X_t; \theta), \tag{2.9}$$

where

$$\Delta^{k+1} R(\Delta, X_t; \theta) = \int_0^\Delta \int_0^{u_1} \cdots \int_0^{u_k} E_\theta(L_\theta^{k+1} f(X_{t+u_{k+1}}) \,|\, X_t) du_{k+1} \cdots du_1,$$

see e.g. Sørensen (2012). The properties of the remainder term follow from Lemma 5.1 in Section 5. When we apply the generator to a function $h(y,x)$ of two variables, we mean

$$L_\theta(h)(y,x) = b(y;\alpha)\partial_y h(y,x) + \tfrac{1}{2} v(y;\beta)\partial_y^2 h(y,x), \tag{2.10}$$

and for a function $h(\Delta, y, x; \theta)$ that depends also on Δ and θ, we use the notation

$$L_\theta(h(\Delta;\tilde{\theta}))(y,x) = b(y;\alpha)\partial_y h(\Delta, y, x; \tilde{\theta}) + \tfrac{1}{2} v(y;\beta)\partial_y^2 h(\Delta, y, x; \tilde{\theta}).$$

The following lemma provides identities that play an essential role in the proofs of the asymptotic theory in the next section. Note that L_θ is applied coordinate-wise to a vector valued function, and that, depending on the context, 0 can also denote a 0-vector.

Lemma 2.4 *Let G_n be an estimating function of the form (1.2), where $g_i \in C_{p,\kappa-1,2(\kappa-1),0}(\mathbb{R} \times (\ell,r)^2 \times \Theta)$, i=1,2, for a $\kappa \geq 2$, and assume Condition 2.2.*

Then G_n is an approximate martingale estimating function of order $\kappa \geq 2$ (i.e. it satisfies (2.7)) if and only if

$$\sum_{i=0}^{k} \binom{k}{i} L_\theta^{k-i}(g^{(i)}(\theta))(x,x) = 0, \quad k = 0, \ldots, \kappa - 1,$$

for all $x \in (\ell, r)$ and $\theta \in \Theta$ (the $g^{(i)}$s are the functions in the expansion (2.1)).

In particular, if G is an approximate martingale estimating function, then

$$g^{(0)}(x,x;\theta) = 0 \tag{2.11}$$
$$g^{(1)}(x,x;\theta) = -L_\theta(g^{(0)}(\theta))(x,x) \tag{2.12}$$

for all $x \in (\ell, r)$ and $\theta \in \Theta$.

2.1 Examples

The prototype of estimating functions satisfying the condition (2.7) are the *martingale estimating functions* for which

$$E_\theta(g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) \,|\, X_{t_{i-1}^n}) = 0.$$

They often have the form

$$g(\Delta, y, x; \theta) = A(x, \Delta; \theta) \left[f(y;\theta) - \pi_\theta^{1,\Delta} f(x;\theta) \right], \tag{2.13}$$

where $f(y;\theta) = (f_1(y;\theta), \ldots, f_N(y;\theta))^T$, with f_i real-valued, $A(x,\Delta;\theta)$ a $2 \times N$-matrix of weights, and $\pi_\theta^{1,\Delta}$ denotes the transition operator given by

$$\pi_\theta^{1,\Delta} h(x) = E_\theta(h(X_\Delta) \,|\, X_0 = x) \tag{2.14}$$

for a real-valued function h. Here and later x^T denotes the transpose of a vector or matrix x. The weight matrix A can be chosen suitably, for instance to obtain rate optimality and efficiency. Examples are polynomial estimating functions, where the real functions f_j are power functions or more general polynomials. The *quadratic martingale estimating function*, obtained for $N = 2$, $f_1(x) = x$ and $f_2(x) = x^2$, is as a useful simple example, see Section 4. Polynomial estimating functions are particularly useful for the class of Pearson diffusions, for which all (finite) moments (conditional as well as unconditional) can be found explicitly, see Forman and Sørensen (2008). Other instances are the estimating functions based on eigenfunctions of the generator (2.8) proposed by Kessler and Sørensen (1999).

The econometric *generalized method of moments* (GMM) based on conditional moments is covered by our theory. This method is in practice often implemented as follows; see Campbell *et al.* (1997). The starting point is an N-dimensional function $h(\Delta, y, x; \theta)$ for which each coordinate satisfies that $E_\theta(h_j(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) \,|\, X_{t_{i-1}^n}) = 0$. Let A_n be an $N \times N$-matrix such that $m_n(\theta) = A_n \sum_{i=1}^n h(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta)$ converges in probability. For the usual low frequency asymptotics, where Δ_n does not depend on n, $A_n = n^{-1} I_N$ (I_N denotes the identity matrix), but for the high frequency asymptotics considered in this paper, a different choice of A_n is usually necessary, as will become clear in the next section. The GMM-estimator is obtained by minimizing $Q_n(\theta) = m_n(\theta)^T W_n \, m_n(\theta)$, where W_n is an $N \times N$-matrix such that $W_n \to W$ in probability. It is typically the (suitably normalized) inverse of a consistent estimator of the covariance matrix of $m_n(\theta)$. Under weak regularity conditions, the GMM-estimator solves the estimating equation $\partial_\theta Q_n(\theta) = \partial_\theta m_n(\theta)^T W_n \, m_n(\theta) = 0$, so if $\partial_\theta m_n(\theta) \to D(\theta)$ in probability (a necessary condition for asymptotic results about the GMM-estimator), then the GMM-estimator has the same asymptotic behavior as the estimator obtained from $D(\theta)^T W A_n \sum_{i=1}^n h(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta)$, which is a martingale estimating function. The close relationship between martingale estimating functions and the type of GMM-estimators described here is discussed in detail in Christensen and Sørensen (2008). More general GMM-estimators of the martingale estimating function type were considered in Hansen (1985, 1993), and a discussion of links between the literature on estimating functions and that on GMM-estimators can be found in Hansen (2001).

Approximate martingale estimating functions can be obtained by replacing the exact conditional expectation in (2.13) by the approximation given by (2.9) such that the function g has the form

$$g(\Delta, y, x; \theta) = A(x, \Delta; \theta) \left[f(y; \theta) - \pi_\theta^{\kappa,\Delta} f(x; \theta) \right], \tag{2.15}$$

where

$$\pi_\theta^{\kappa,\Delta} f(x; \theta) = \sum_{i=0}^{\kappa-1} \frac{\Delta^i}{i!} \, L_\theta^i f(x; \theta), \qquad \kappa = 2, 3, \ldots, \tag{2.16}$$

with the generator L_θ applied coordinate-wise. This estimating functions satisfies (2.7). A simple example is $g(\Delta, y, x; \theta) = a(x, \Delta; \theta)(y - x - b(x; \alpha)\Delta)$ with $\kappa = 2$, considered by Prakasa Rao (1988) and Florens-Zmirou (1989). Other instances are the estimators proposed by Chan *et al.* (1992) and Kelly *et al.* (2004). For all $\kappa \in \mathbb{N}$, ($\kappa \geq 2$), Kessler (1997) proposed a *Gaussian approximation* to the likelihood function, for which the corresponding pseudo-score function is an approximate martingale estimating function that satisfies (2.7).

3 Optimal rate

In this section we present asymptotic results for approximate martingale estimating functions. We begin with a general approximate martingale estimating function. Then we will see how Condition 1.1 implies rate optimality, so that the estimator of the parameter in the diffusion coefficient converges faster than the estimator of the parameter in the drift coefficient. As previously, x^T denotes the transpose of a vector or matrix x and $\theta_0 = (\alpha_0, \beta_0)$ denotes the true parameter value.

Theorem 3.1 *Assume that the Conditions 2.2 and 2.3 hold. Suppose, moreover, the identifiability condition that*

$$\gamma(\theta,\theta_0) = \int_\ell^r [b(x,\alpha_0) - b(x,\alpha)]\partial_y g(0,x,x;\theta)\mu_{\theta_0}(x)dx \\ + \tfrac{1}{2}\int_\ell^r [v(x,\beta_0) - v(x,\beta)]\partial_y^2 g(0,x,x;\theta)\mu_{\theta_0}(x)dx \neq 0 \tag{3.1}$$

for all $\theta \neq \theta_0$, and that the matrix

$$S = \int_\ell^r J_{\theta_0}(x)\mu_{\theta_0}(x)dx \tag{3.2}$$

is invertible, where

$$J_\theta(x) = \begin{pmatrix} \partial_\alpha b(x;\alpha)\partial_y g_1(0,x,x;\theta) & \frac{1}{2}\partial_\beta v(x;\beta)\partial_y^2 g_1(0,x,x;\theta) \\ \partial_\alpha b(x;\alpha)\partial_y g_2(0,x,x;\theta) & \frac{1}{2}\partial_\beta v(x;\beta)\partial_y^2 g_2(0,x,x;\theta) \end{pmatrix}. \tag{3.3}$$

Then with a probability that goes to one as $n \to \infty$, a consistent G_n-estimator $\hat\theta_n = (\hat\alpha_n, \hat\beta_n)$ exists and is unique in any compact subset K of Θ with $\theta_0 \in \mathrm{int}\,K$. If $n\Delta_n^{2\kappa-1} \to 0$, then

$$\sqrt{n\Delta_n}(\hat\theta_n - \theta_0) \xrightarrow{\mathcal{D}} N_2\left(0, S^{-1}V_0(S^T)^{-1}\right) \tag{3.4}$$

under P_{θ_0}, where $V_0 = V(\theta_0)$ with

$$V(\theta) = \int_\ell^r v(x,\beta_0)\partial_y g(0,x,x;\theta)\partial_y g(0,x,x;\theta)^T\mu_{\theta_0}(x)dx. \tag{3.5}$$

For a martingale estimating function (3.4) holds without the extra condition on the rate of convergence of Δ_n.

A consistent estimator of the asymptotic variance of $\hat\theta_n$ can be obtained from

$$\frac{1}{n\Delta_n}\sum_{i=1}^n \partial_{\theta^T} g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\hat\theta_n) \xrightarrow{P_{\theta_0}} -S \tag{3.6}$$

$$\frac{1}{n\Delta_n}\sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\hat\theta_n) g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\hat\theta_n)^T \xrightarrow{P_{\theta_0}} V_0 \tag{3.7}$$

The theorem follows from general asymptotic statistical results for stochastic processes, see e.g. Jacod and Sørensen (2018). The proof is given in Section 5. The precise meaning of the uniqueness statement is that for any G_n-estimator $\tilde{\theta}_n$ with $P_{\theta_0}(\tilde{\theta}_n \in K) \to 1$ as $n \to \infty$, it holds that $P_{\theta_0}(\hat{\theta}_n \neq \tilde{\theta}_n) \to 0$ as $n \to \infty$.

We see from (3.4) that the rate of convergence of both $\hat{\alpha}$ and $\hat{\beta}$ is $1/\sqrt{n\Delta_n}$, provided that the matrix V_0 is regular. Here $n\Delta_n$ is the length of the interval in which the diffusion is observed. Gobet (2002) showed that under weak regularity conditions a discretely sampled diffusion model is local asymptotically normal in the high frequency/infinite time horizon asymptotic scenario considered here, and that the optimal rate of convergence for estimators of parameters in the drift coefficient is $1/\sqrt{n\Delta_n}$, whereas the optimal rate for estimators of parameters in the diffusion coefficient is $1/\sqrt{n}$.

The next theorem shows what happens when Jacobsen's condition, Condition 1.1, is satisfied, or more precisely, when a version of the estimating function satisfies the condition.

Theorem 3.2 *Suppose the Conditions 1.1, 2.2 and 2.3 hold. Assume, moreover, that the following identifiability condition is satisfied*

$$\begin{aligned} \int_\ell^r [b(x,\alpha_0) - b(x,\alpha)]\partial_y g_1(0,x,x;\theta)\mu_{\theta_0}(x)dx &\neq 0 \qquad \textit{when } \alpha \neq \alpha_0 \\ \int_\ell^r [v(x,\beta_0) - v(x,\beta)]\partial_y^2 g_2(0,x,x;\theta)\mu_{\theta_0}(x)dx &\neq 0 \qquad \textit{when } \beta \neq \beta_0, \end{aligned}$$

and that $S_{11} \neq 0$ and $S_{22} \neq 0$ with S given by (3.2). Then with a probability that goes to one as $n \to \infty$, a consistent G_n-estimator $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)$ exists and is unique in any compact subset K of Θ with $\theta_0 \in int K$.

If, moreover,

$$\partial_\alpha \partial_y^2 g_2(0,x,x;\theta) = 0 \tag{3.8}$$

and $n\Delta_n^{2(\kappa-1)} \to 0$, then

$$\begin{pmatrix} \sqrt{n\Delta_n}(\hat{\alpha}_n - \alpha_0) \\ \sqrt{n}(\hat{\beta}_n - \beta_0) \end{pmatrix} \xrightarrow{\mathcal{D}} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} W_1(\theta_0)/S_{11}^2 & 0 \\ 0 & W_2(\theta_0)/S_{22}^2 \end{pmatrix} \right) \tag{3.9}$$

where

$$\begin{aligned} W_1(\theta) &= \int_\ell^r v(x;\beta_0)[\partial_y g_1(0,x,x;\theta)]^2 \mu_{\theta_0}(x)dx = V(\theta)_{11} \\ W_2(\theta) &= \tfrac{1}{2}\int_\ell^r [v(x;\beta_0)^2 + \frac{1}{2}(v(x;\beta_0) - v(x;\beta))^2][\partial_y^2 g_2(0,x,x;\theta)]^2 \mu_{\theta_0}(x)dx \end{aligned}$$

with $V(\theta)$ given by (3.5). For a martingale estimating function (3.9) holds without the extra condition on the rate of convergence of Δ_n.

A consistent estimator of the asymptotic variance of $\hat{\theta}_n$ can be obtained from (3.6) and

$$D_n \sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \hat{\theta}_n) g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \hat{\theta}_n)^T D_n \xrightarrow{P_{\theta_0}} \begin{pmatrix} W_1(\theta_0) & 0 \\ 0 & W_2(\theta_0) \end{pmatrix}, \tag{3.10}$$

where

$$D_n = \begin{pmatrix} \frac{1}{\sqrt{n\Delta_n}} & 0 \\ 0 & \frac{1}{\Delta_n\sqrt{n}} \end{pmatrix}. \tag{3.11}$$

Thus Jacobsen's condition (1.3) and the additional condition (3.8) imply rate optimal estimators and that the estimators of the drift parameter and of the diffusion coefficient parameter are asymptotically independent. In the next section we shall see that (3.8) is automatically satisfied for efficient estimating functions. Note that for non-martingale estimating functions Δ_n must go a bit faster to zero than was required in Theorem 3.1. Note also that if the first coordinate of g satisfies Jacobsen's condition too, then the first part of the identifiability condition in Theorem 3.2 does not hold, and the parameter α cannot be consistently estimated by the estimating function (1.2). The proof of Theorem 3.2 is given in Section 5.

Example 3.3 Consider a *quadratic martingale estimating function* of the form

$$g(\Delta, y, x; \theta) = \begin{pmatrix} a_1(x, \Delta; \theta)[y - F(\Delta, x; \theta)] \\ a_2(x, \Delta; \theta)\left[(y - F(\Delta, x; \theta))^2 - \phi(\Delta, x; \theta)\right] \end{pmatrix}, \tag{3.12}$$

where $F(\Delta, x; \theta) = E_\theta(X_\Delta | X_0 = x)$ and $\phi(\Delta, x; \theta) = \mathrm{Var}_\theta(X_\Delta | X_0 = x)$. Since, by (2.9), $F(\Delta, x; \theta) = x + O(\Delta)$ and $\phi(\Delta, x; \theta) = O(\Delta)$, we find that

$$g(0, y, x; \theta) = \begin{pmatrix} a_1(x, 0; \theta)(y - x) \\ a_2(x, 0; \theta)(y - x)^2 \end{pmatrix}. \tag{3.13}$$

Jacobsen's condition (1.3) is satisfied because $\partial_y g_2(0, y, x; \theta) = 2a_2(x, \Delta; \theta)(y - x)$. Thus estimators obtained from (3.12) are rate optimal, provided that (3.8) is satisfied, for instance if a_2 does not depend on α.

Clearly (3.13) holds if F and ϕ in (3.12) are replaced by expansions of order $O(\Delta^{\kappa-1})$ with $\kappa \geq 2$, using again (2.9). Thus rate optimal estimators are also obtained in this more easily calculated case, provided again that (3.8) holds. The simplest example ($\kappa = 2$) is

$$g(\Delta, y, x; \theta) = \begin{pmatrix} a_1(x, \Delta; \theta)[y - x - b(x; \alpha)\Delta] \\ a_2(x, \Delta; \theta)\left[(y - x - b(x; \alpha)\Delta)^2 - v(\Delta, x; \beta)\Delta\right] \end{pmatrix}. \tag{3.14}$$

It is instructive to consider an example of an estimating function for which estimators are not rate optimal. The martingale estimating function

$$g(\Delta, y, x; \theta) = \begin{pmatrix} a_1(x, \Delta; \theta)[y - F(\Delta, x; \theta)] \\ a_2(x, \Delta; \theta)\left[y^2 - (\phi(\Delta, x; \theta) + F(\Delta, x; \theta)^2)\right] \end{pmatrix} \tag{3.15}$$

does not satisfy (1.3). It is easy to check that a version of g satisfying (1.3) exists if and only if $a_1(x, 0; \theta) = c_\theta a_2(x, 0; \theta)x$ for some real constant c_θ. In all other cases, the estimating function given by (3.15) is not rate optimal. We can obtain a particular case of (3.12) from (3.15) by choosing $a_1(x, \Delta; \theta) = a_2(x, \Delta; \theta)F(\Delta, x; \theta)$ and using the version $\tilde{g}_1 = g_1$, $\tilde{g}_2 = g_2 - 2g_1$. This particular case of (3.15), obviously satisfies that $a_1(x, 0; \theta) = a_2(x, 0; \theta)x$.

□

4 Efficient estimating functions

In this section we study the conditions under which an approximate martingale estimating function, $G_n(\theta)$, yields an efficient estimator. In particular, we show that Condition 1.2 ensures efficiency, and that Godambe-Heyde optimal estimating functions yield rate optimal and efficient estimators.

Theorem 4.1 *Suppose Condition 1.2 and the conditions of Theorem 3.2 except (3.8) are satisfied. Then the conclusions of Theorem 3.2 hold and the estimating function (1.2) is efficient, i.e., the asymptotic covariance matrix of the estimator* $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)$ *equals*

$$\Sigma(\theta_0) = \begin{pmatrix} \left(\int_\ell^r \frac{(\partial_\alpha b(x;\alpha_0))^2}{v(x;\beta_0)} \mu_{\theta_0}(x)dx\right)^{-1} & 0 \\ 0 & 2\left(\int_\ell^r \left[\frac{\partial_\beta v(x;\beta_0)}{v(x;\beta_0)}\right]^2 \mu_{\theta_0}(x)dx\right)^{-1} \end{pmatrix}. \tag{4.1}$$

Consistent estimators of the asymptotic variances can be obtained from

$$\frac{1}{n\Delta_n}\sum_{i=1}^n g_1(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \hat{\theta}_n)^2 \xrightarrow{P_{\theta_0}} \int_\ell^r \frac{(\partial_\alpha b(x;\alpha_0))^2}{v(x;\beta_0)} \mu_{\theta_0}(x)dx$$

and

$$\frac{1}{n\Delta_n^2}\sum_{i=1}^n g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \hat{\theta}_n)^2 \xrightarrow{P_{\theta_0}} \int_\ell^r \left[\frac{\partial_\beta v(x;\beta_0)}{v(x;\beta_0)}\right]^2 \mu_{\theta_0}(x)dx.$$

Note that an efficient estimating function automatically satisfies (3.8). An asymptotic martingale estimating function is efficient if and only if there exists a version that satisfies the conditions of Theorem 4.1.

The covariance matrix (4.1) is equal to the leading term in the expansion of the asymptotic variance of the maximum likelihood estimator in powers of Δ found by Dacunha-Castelle and Florens-Zmirou (1986). The asymptotic variance of $\hat{\alpha}_n$ equals that of the maximum likelihood estimator based on continuous time observation, see e.g. Kutoyants (2004).

Example 4.2 Consider again the *quadratic martingale estimating function* (3.12). The function $g(0, y, x; \theta)$, given by (3.13), satisfies the conditions for efficiency (1.4) and (1.5) if we choose $a_1(x, \Delta; \theta) = \partial_\alpha b(x;\alpha)/\sigma^2(x;\beta)$ and $a_2(x, \Delta; \theta) = \partial_\beta \sigma^2(x;\beta)/\sigma^4(x;\beta)$, as proposed by Bibby and Sørensen (1995, 1996). The same is true of weight functions a_1 and a_2 that converge to $\partial_\alpha b/\sigma^2$ and $\partial_\beta \sigma^2/\sigma^4$ as $\Delta \to 0$. An example is $a_1(x, \Delta; \theta) = \partial_\alpha F(\Delta, x; \theta)/\phi(\Delta, x; \theta)$ and $a_2(x, \Delta; \theta) = \Delta \partial_\beta \phi(\Delta, x; \theta)/\phi(\Delta, x; \theta)^2$. This is the optimal quadratic martingale estimating function in the sense of Godambe and Heyde (1987) (after multiplication of the second coordinate by Δ), see Bibby and Sørensen (1995, 1996).

Consider the pseudo-likelihood function obtained from the likelihood function by replacing the transition density $p(\Delta, y, x; \theta)$ by the Gaussian density with mean $F(\Delta, x; \theta)$ and variance $\phi(\Delta, x; \theta)$. The exact conditional moments are used to ensure consistency of the estimator also in the case of low frequency asymptotics, where Δ is not small. The corresponding pseudo-score function is the quadratic estimating function (3.12) with

$a_1(x,\Delta;\theta) = \partial_\alpha F(\Delta,x;\theta)/\phi(\Delta,x;\theta)$ and $a_2(x,\Delta;\theta) = \partial_\beta \phi(\Delta,x;\theta)/\phi(\Delta,x;\theta)^2$, which we have just seen is efficient.

A pseudo-likelihood function that works for data sampled at a high frequency is the likelihood function obtained by replacing the original diffusion model by its Euler approximation. It can be obtained from the original likelihood function by replacing the transition density by the Gaussian density with mean and variance given by the expansions $x - b(x;\alpha)\Delta$ and $\sigma^2(x;\beta)\Delta$. The corresponding pseudo score is of the form (3.14) with $a_1(x,\Delta;\theta) = \partial_\alpha b(x;\alpha)/\sigma^2(x;\beta)$ and (after multiplication by Δ) $a_2(x,\Delta;\theta) = \partial_\beta \sigma^2(x;\beta)/\sigma^4(x;\beta)$. Since (3.13) holds, this Euler pseudo score function satisfies the conditions for efficiency. This estimator has often been used in empirical work in finance. Similarly, it follows that the estimators considered by Dorogovcev (1976), Prakasa Rao (1988), Florens-Zmirou (1989), Yoshida (1992), Kessler (1997), Kelly *et al.* (2004), and Uchida and Yoshida (2013) are efficient under suitable conditions on the rate of convergence of Δ_n.

□

Example 4.3 A final example is *maximum likelihood estimation.* In broad generality, the score function is a martingale estimating function, see e.g. Barndorff-Nielsen and Sørensen (1994). The transition density can, under weak regularity conditions, be expanded in powers of Δ

$$p(\Delta,y,x;\theta) = r(\Delta,y,x;\theta)(1+O(\Delta)),$$

where

$$r(\Delta,y,x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2(y;\beta)\Delta}} \exp\left(-\frac{(k(y;\beta)-k(x;\beta))^2}{2\Delta} + m(y;\alpha,\beta) - m(x;\alpha,\beta) - \tfrac{1}{2}\log\left(\frac{\sigma(y;\beta)}{\sigma(x;\beta)}\right)\right),$$

$k(x;\beta) = \int^x \sigma^{-1}(z;\beta)dz$ and $m(x;\alpha,\beta) = \int^x b(z;\alpha)/\sigma^2(z;\beta)dz$, see e.g. Dacunha-Castelle and Florens-Zmirou (1986) or Gihman and Skorohod (1972), Part I, Chapter 13. Therefore, under regularity conditions, the (suitably normalized) score function $g_1(\Delta,y,x;\theta) = \partial_\alpha \log p(\Delta,y,x;\theta)$ and $g_2(\Delta,y,x;\theta) = \Delta\partial_\beta \log p(\Delta,y,x;\theta)$ satisfies that

$$\begin{aligned} g_1(\Delta,y,x;\theta) &= \int_x^y \frac{\partial_\alpha b(z;\alpha)}{\sigma^2(z;\beta)}dz + O(\Delta) \\ g_2(\Delta,y,x;\theta) &= -[k(y;\beta)-k(x;\beta)][\partial_\beta k(y;\beta) - \partial_\beta k(x;\beta)] + O(\Delta). \end{aligned}$$

From these expansions it follows that the score functions (normalized as above) satisfies the conditions (1.3), (1.4) and (1.5) for rate optimality and efficiency. In particular, $\partial_y^2 g_2(0,x,x;\theta) = -2\partial_x k(x;\beta)\partial_\beta\partial_x k(x;\beta) = \partial_\beta\sigma^2(x;\beta)/\sigma^4(x;\beta)$. Obviously, also the pseudo-likelihood function obtained by replacing the transition density p by r has a pseudo-score function that satisfies the conditions for rate optimality and efficiency.

□

Consider martingale estimating functions of the form (2.13) and the related approximate martingale estimating functions (2.15), i.e.

$$G_n(\theta) = \sum_{i=1}^n A(X_{t_{i-1}^n},\Delta;\theta)[f(X_{t_i^n};\theta) - \pi_\theta^{\kappa,\Delta} f(X_{t_{i-1}^n};\theta)], \quad \kappa = 1,2,\ldots, \tag{4.2}$$

with $\pi_\theta^{\kappa,\Delta}$ given by (2.14) for $\kappa = 1$ (the martingale case) and by (2.16) for $\kappa = 2, 3, \ldots$. Moreover, A is a $2 \times N$-matrix of weights, and we assume that the coordinates of the N-dimensional function $f(x, \theta)$ are twice continuously differentiable w.r.t. x.

For estimating functions of the form (4.2), the condition for rate optimality is

$$\sum_{j=1}^{N} a_{2j}(x, 0; \theta) \partial_x f_j(x; \theta) = 0, \tag{4.3}$$

where a_{ij} denotes the ijth entry of A, and the condition for efficiency is

$$\begin{aligned} \sum_{j=1}^{N} a_{1j}(x, 0; \theta) \partial_x f_j(x; \theta) &= \partial_\alpha b(x; \alpha)/\sigma^2(x; \beta) & (4.4) \\ \sum_{j=1}^{N} a_{2j}(x, 0; \theta) \partial_x^2 f_j(x; \theta) &= \partial_\beta \sigma^2(x; \beta)/\sigma^4(x; \beta) & (4.5) \end{aligned}$$

For a given function f, we want to find a weight-matrix A such that these equations are satisfied. Obviously, it is necessary that $N \geq 2$ in order that all three equations are satisfies. If $N = 1$, an efficient approximate martingale estimating function can be obtained by solving (4.4), provided that the diffusion coefficient is known, so that only the drift depends on a parameter.

First consider $N = 2$, and assume that the matrix

$$M(x) = \begin{pmatrix} \partial_x f_1(x; \theta) & \partial_x^2 f_1(x; \theta) \\ \partial_x f_2(x; \theta) & \partial_x^2 f_2(x; \theta) \end{pmatrix} \tag{4.6}$$

is invertible for μ_θ-almost all x. Then the linear equations (4.3) - (4.5) are satisfied for

$$A(x, 0; \theta) = \begin{pmatrix} \partial_\alpha b(x; \alpha)/\sigma^2(x; \beta) & c(x; \theta) \\ 0 & \partial_\beta \sigma^2(x; \beta)/\sigma^4(x; \beta) \end{pmatrix} M(x)^{-1}, \tag{4.7}$$

where $c(x; \theta)$ is any (measurable) function. As a simple example, the quadratic estimating function ($N = 2$, $f_1(x) = x$ and $f_2(x) = \frac{1}{2}x^2$) is rate optimal and efficient if the weights are weights $a_{11}(x) = \partial_\alpha b(x; \alpha)/\sigma^2(x; \beta)$, $a_{12}(x) = 0$, $a_{22}(x) = \partial_\beta \sigma^2(x; \beta)/\sigma^4(x; \beta)$ and $a_{21}(x) = -2x a_{22}(x)$ (we have chosen $c = 0$). Note that a simple choice for the weight matrix $A(x, \Delta; \theta)$ for $\Delta > 0$ is to let it be given by (4.7) for all Δ.

It is easily seen that, for any $N \geq 2$, there exist many solutions to (4.3) - (4.5) provided that there are two coordinates of f (without loss of generality, we can assume these to be f_1 and f_2) such that $M(x)$ is invertible. In the special case $\kappa = 1$, this result follows from Theorem 2.2 of Jacobsen (2002). The conditions for Jacobsen's concept small Δ-optimality for martingale estimating functions are identical to our conditions for rate optimality and efficiency, so we can take advantage of his thorough study of when martingale estimating functions of the type (4.2) with $\kappa = 1$ are small Δ-optimal.

Here we will, however, go another way and give a natural and generally useful way of finding a rate optimal and efficient weight matrix A for all $\kappa \geq 1$. A weight matrix A^* in a martingale estimating function of the type (4.2) (i.e. $\kappa = 1$) is optimal in the sense of Godambe and Heyde (1987), see also Heyde (1997), if it solves the linear equation

$$A^*(x,\Delta;\theta)\,E_\theta\left([f(X_\Delta;\theta)-\pi_\theta^{1,\Delta}f(x;\theta)][f(X_\Delta;\theta)-\pi_\theta^{1,\Delta}f(x;\theta)]^T\,|\,X_0=x\right) \tag{4.8}$$
$$=\partial_\theta\pi_\theta^{1,\Delta}f^T(x;\theta)-\pi_\theta^{1,\Delta}\partial_\theta f^T(x;\theta).$$

It can be assumed that the functions $f_1,\ldots,f_N$ are affinely independent such that the conditional covariance matrix in (4.8) is invertible. The Godambe-Heyde optimal martingale estimating function gives an estimator that minimizes the asymptotic variance of estimators obtained from the class of martingale estimating functions of the form (4.2) (with $\kappa = 1$) with a fixed function f and for a fixed, possibly large, Δ. The next theorem shows that the Godambe-Heyde optimal estimators are rate optimal and efficient in the high frequency asymptotic scenario considered in the present paper. Moreover, the same is true of the approximate martingale estimating functions obtained by expanding all conditional moments (including those in A^*) in powers of Δ, which gives a feasible general way of constructing explicit estimating functions that are rate optimal and efficient.

Theorem 4.4 *Suppose Condition 2.2 is satisfied, that $f_j \in C_{p,6,1}((\ell,r)\times\Theta)$, $j=1,\ldots,N$, that $N\geq 2$ and that the 2×2 matrix $M(x)$ given by (4.6) is invertible for μ_θ-almost all x. Let $A^*(x,\Delta;\theta)$ satisfy (4.8), and define*

$$B(x,\Delta,,\theta)=\begin{pmatrix}1 & 0\\ 0 & 2\Delta\end{pmatrix}A^*(x,\Delta;\theta).$$

Then the limit $B(x,0,\theta)$ exists, and

$$g^*(\Delta,y,x;\theta)=B(x,\Delta,,\theta)[f(y;\theta)-\pi_\theta^{\kappa,\Delta}f(x;\theta)] \tag{4.9}$$

satisfies the conditions for rate optimality (1.3) and efficiency (1.4) and (1.5) for all $\kappa\in\mathbb{N}$. The same is true if B is replaced by a matrix $\tilde{B}$ satisfying that $\tilde{B}(x,0,\theta)=B(x,0,\theta)$.

The matrix $\tilde{B}$ can be obtained by replacing the conditional moments in A^* by expansions in powers of Δ, or simply by defining $\tilde{B}(x,\Delta,\theta)=B(x,0,\theta)$. It is surprising that a local property like Godambe-Heyde optimality, which ensures optimality only within a particular class of estimating functions, implies global optimality properties like rate optimality and efficiency. Phrased in terms of the concept small Δ-optimality, this result was conjectured by Jacobsen (2002) for martingale estimating functions ($\kappa=1$). The fact that one of the conditions for efficiency is $N\geq 2$ explains the finding in Larsen and Sørensen (2007) that an optimal martingale estimating function based on two eigenfunctions seemed to be efficient for weekly observations of exchange rates in a target zone.

Let us conclude this section by stating the results for a d-dimensional diffusion. In this case $b(x;\alpha)$ is d-dimensional and $v(x;\beta)=\sigma(x;\beta)\sigma(x;\beta)^T$ is a $d\times d$-matrix. The conditions for efficiency are

$$\partial_y g_1(0,x,x;\theta)=\partial_\alpha b(x;\alpha)^T v(x;\beta)^{-1}$$

and

$$\mathrm{vec}\left(\partial_y^2 g_2(0,x,x;\theta)\right)=\mathrm{vec}\left(\partial_\beta v(x;\beta)\right)\left(v^{\otimes 2}(x;\beta)\right)^{-1}.$$

In the latter equation, $\mathrm{vec}(M)$ denotes for a $d \times d$ matrix M the d^2-dimensional row vector consisting of the rows of M placed one after the other, and $M^{\otimes 2}$ is the $d^2 \times d^2$-matrix with $(i', j'), (ij)$th entry equal to $M_{i'i}M_{j'j}$. Thus if $M = \partial_\beta v(x;\beta)$ and $M^\bullet = (v^{\otimes 2}(x;\beta))^{-1}$, then the (i,j)th coordinate of $\mathrm{vec}(M)\, M^\bullet$ is $\sum_{i'j'} M_{i'j'} M^\bullet_{(i'j'),(i,j)}$. These expressions are the conditions for small Δ-optimality for multivariate diffusions given by Jacobsen (2002).

For a d-dimensional diffusion process, the condition analogous to the one discussed shortly before Theorem 4.4 ensuring the existence of a rate optimal and efficient estimating function of the form (4.2) is that $N \geq d(d+3)/2$, and that the $N \times (d + d^2)$-matrix

$$\left(\begin{array}{cc} \partial_x f(x;\theta) & \partial_x^2 f(x;\theta) \end{array} \right)$$

has full rank $d(d+3)/2$. For $\kappa = 1$ this follows from Theorem 2.2 of Jacobsen (2002), and it is clear from the proof of this theorem that it holds for $\kappa \geq 2$ too. When α and β are multivariate, we further need that $\{\partial_{\alpha_i} b(x;\alpha)\}$ and $\{\partial_{\beta_i} v(x;\beta)\}$ are two sets of linearly independent functions of x. These conditions also ensure that Theorem 4.4 holds for a d-dimensional diffusion process, i.e. that the Godambe-Heyde optimal martingale estimating function is rate optimal and efficient for a d-dimensional diffusion process.

5 Proofs and lemmas

The first of the following lemmas is a slight generalization of Lemma 6 in Kessler (1997), while the second lemma is essentially Lemma 8 in the same paper. The proofs are analogous to those in Kessler's paper. The result (5.3) follows from (5.2). The notation $R(\Delta, y, x; \theta)$ was defined in Section 2. We sometimes use the notation $a \leq_C b$, which means that there exists a $C > 0$ such that $a \leq Cb$.

Lemma 5.1 *Assume Condition 2.2. Then a constant $C_k > 0$ exists for $k = 1, 2, \ldots$ such that*

$$E_{\theta_0}(|X_{t+\Delta} - X_t|^k \mid X_t) \leq C_k \Delta^{k/2}(1 + |X_t|)^{C_k} \tag{5.1}$$

for $\Delta > 0$. Let $f(y, x, \theta)$ be a real function of polynomial growth in x and y uniformly for θ in a compact set K. Then for any fixed $\Delta_0 > 0$ there exists a constant $C > 0$ such that

$$E_{\theta_0}(|f(X_{t+\Delta}, X_t, \theta)| \mid X_t) \leq C(1 + |X_t|)^C \quad \text{for } \Delta \in [0, \Delta_0] \text{ and } \theta \in K. \tag{5.2}$$

Suppose the function $f(y, x, \theta)$ is, moreover, $2k$ times differentiable ($k \leq 3$) with respect to y with all derivatives of polynomial growth in x and y uniformly for θ in compact sets. Then

$$\int_0^\Delta \int_0^{u_1} \cdots \int_0^{u_{k-1}} E_{\theta_0}\left(L_{\theta_0}^k(f)(X_{t+u_k}, X_t; \theta) \mid X_t \right) du_k \cdots du_1 = \Delta^k R(\Delta, X_t, \theta). \tag{5.3}$$

The result (5.3) is used to ensure that the remainder term in expansions of the type (2.9) have the expected order. The result can be proved for larger values of k if stronger differentiability conditions are imposed on the coefficients b and σ.

Lemma 5.2 *Assume Condition 2.2, and let $f(x,\theta)$ be a real function that is differentiable with respect to x and θ with derivatives of polynomial growth in x uniformly for θ in a compact set. Then*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_{t_i^n};\theta) \xrightarrow{P_{\theta_0}} \int_\ell^r f(x;\theta)\mu_{\theta_0}(x)dx$$

uniformly for θ in a compact set.

Lemma 9 in Genon-Catalot and Jacod (1993) is used frequently in the proofs of Lemma 5.5 and Lemma 5.6 to establish pointwise convergence. The result is therefore cited here for the convenience of the reader.

Lemma 5.3 *Let Z_i^n ($i = 1,\ldots,n, n \in \mathbb{N}$) be a triangular array of random variables such that Z_i^n is $\mathcal{G}_i^n$-measurable, where $\mathcal{G}_i^n = \sigma(W_s : s \le t_i^n)$. If*

$$\sum_{i=1}^{n} E_\theta(Z_i^n \,|\, \mathcal{G}_{i-1}^n) \xrightarrow{P_\theta} U$$

and

$$\sum_{i=1}^{n} E_\theta((Z_i^n)^2 \,|\, \mathcal{G}_{i-1}^n) \xrightarrow{P_\theta} 0,$$

where U is a random variable, then

$$\sum_{i=1}^{n} Z_i^n \xrightarrow{P_\theta} U.$$

Proof of Lemma 2.4. Combining (2.1) and (2.9), we find that

$$E_\theta(g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \,|\, X_{t_{i-1}^n}) = \sum_{\ell=0}^{\kappa-1} \frac{\Delta_n^\ell}{\ell!} \sum_{j=0}^{\ell} \binom{\ell}{j} L_\theta^{\ell-j}(g^{(j)}(\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n}) + \Delta_n^\kappa R(\Delta, X_{t_{i-1}^n}, \theta),$$

from which the "if" statement of the lemma follows immediately. The "only if" statement follows from the same expansion because an approximate martingale estimating function satisfies $E_\theta(g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \,|\, X_{t_{i-1}^n}) = O(\Delta_n^\kappa)$.

□

Theorem 3.1 follows via asymptotic statistical results for stochastic processes, see e.g. Jacod and Sørensen (2018). To prove the theorem we need two technical lemmas. The first is used to establish uniform convergence in the proofs of Lemma 5.5 and Lemma 5.6. The lemma is easier to formulate with the following definitions.

Let $\mathcal{C}_0$ denote the subclass of $C_{p,1,2,1}(\mathbb{R}_+, (\ell,r)^2, \Theta)$ of functions $f(\Delta, y, x;\theta)$ satisfying that $f(0,x,x;\theta) = 0$ for all $x \in (\ell, r)$ and $\theta \in \Theta$, and define the operators

$$\begin{aligned}
\mathcal{L}_1 f(s,y,x;\theta) &= \partial_s f(s,y,x;\theta) + \partial_y f(s,y,x;\theta) b(y;\alpha_0) + \tfrac{1}{2}\partial_y^2 f(s,y,x;\theta) v(y,\beta_0) \\
\mathcal{L}_2 f(s,y,x;\theta) &= \partial_y f(s,y,x;\theta)\sigma(y;\beta_0).
\end{aligned}$$

Lemma 5.4 *Assume Condition 2.2, and consider*

$$\zeta_n^{(j)}(\theta) = \frac{1}{n\Delta_n^{j/2}} \sum_{i=1}^n f(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta), \quad j = 2, 3, 4, \tag{5.4}$$

for $f \in \mathcal{C}_0$. Then the following holds for $j = 2$. For every $m \in \mathbb{N}$ and for every compact $K \subseteq \Theta$, a constant $C_{m,K} > 0$ exists such that

$$E_{\theta_0}\left(|\zeta_n^{(j)}(\theta_2) - \zeta_n^{(j)}(\theta_1)|^{2m}\right) \leq C_{m,K}|\theta_2 - \theta_1|^{2m} \tag{5.5}$$

for all θ_1 and θ_2 in K and for all n.

Moreover, if $h_i = \mathcal{L}_i f \in \mathcal{C}_0$ for $i = 1, 2$, then (5.5) holds for $j = 3$, and if $\mathcal{L}_2 h_i \in \mathcal{C}_0$ for $i = 1, 2$, then (5.5) holds for $j = 4$.

Proof. By Ito's formula

$$\begin{aligned} & f(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) = \\ & \int_{t_{i-1}^n}^{t_i^n} h_1(s - t_{i-1}^n, X_s, X_{t_{i-1}^n}; \theta) ds + \int_{t_{i-1}^n}^{t_i^n} h_2(s - t_{i-1}^n, X_s, X_{t_{i-1}^n}; \theta) dW_s. \end{aligned} \tag{5.6}$$

We treat the two terms on the right hand side of (5.6) separately. For a function $k(s, y, x; \theta)$, define $Dk(\cdot; \theta_2, \theta_1) = k(\cdot; \theta_2) - k(\cdot; \theta_1)$. Because $f \in \mathcal{C}_0$, the partial derivatives $\partial_\theta h_i$, $i = 1, 2$, are of polynomial growth in y and x uniformly for θ in a compact set. Therefore

$$\begin{aligned} & \frac{1}{\Delta_n^{2m}} E_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} Dh_1(s - t_{i-1}^n, X_s, X_{t_{i-1}^n}; \theta_2, \theta_1) ds\right|^{2m}\right) \\ \leq \;& \frac{1}{n\Delta_n^{2m}} \sum_{i=1}^n E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{t_i^n} Dh_1(s - t_{i-1}^n, X_s, X_{t_{i-1}^n}; \theta_2, \theta_1) ds\right|^{2m}\right) \\ \leq \;& \frac{1}{n\Delta_n} \sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} E_{\theta_0}\left(|Dh_1(s - t_{i-1}^n, X_s, X_{t_{i-1}^n}; \theta_2, \theta_1)|^{2m}\right) ds \\ \leq_C \;& \frac{1}{n\Delta_n} \sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} E_{\theta_0}\left(\left|\int_0^1 \partial_\theta h_1(s - t_{i-1}^n, X_s, X_{t_{i-1}^n}; \theta_1 + u(\theta_2 - \theta_1)) du\right|^{2m}\right) ds |\theta_2 - \theta_1|^{2m} \\ \leq_C \;& |\theta_2 - \theta_1|^{2m}, \end{aligned}$$

where we have used Condition 2.2 and Jensen's inequality (twice). Using the Burkholder-

Davis-Gundy inequality and Jensen's inequality we obtain

$$\begin{aligned}
&\frac{1}{\Delta_n^{2m}} E_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} Dh_2(s-t_{i-1}^n, X_s, X_{t_{i-1}^n};\theta_2,\theta_1)dW_s\right|^{2m}\right)\\
\le_C\ & \frac{1}{\Delta_n^{2m}} E_{\theta_0}\left(\left|\frac{1}{n^2}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} Dh_2(s-t_{i-1}^n, X_s, X_{t_{i-1}^n};\theta_2,\theta_1)^2 ds\right|^{m}\right)\\
\le\ & \frac{1}{n^{m+1}\Delta_n^{2m}} \sum_{i=1}^n E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{t_i^n} Dh_2(s-t_{i-1}^n, X_s, X_{t_{i-1}^n};\theta_2,\theta_1)^2 ds\right|^{m}\right)\\
\le\ & \frac{1}{(n\Delta_n)^{m+1}} \sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} E_{\theta_0}\left(|Dh_2(s-t_{i-1}^n, X_s, X_{t_{i-1}^n};\theta_2,\theta_1)|^{2m}\right) ds\\
\le_C\ & \frac{1}{(n\Delta_n)^m}|\theta_2-\theta_1|^{2m},
\end{aligned}$$

which implies (5.5) for $j=2$.

The result for $j=3,4$ can be proved in a similar way. When $h_i \in \mathcal{C}_0$ for $i=1,2$,

$$\begin{aligned}
&f(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) = \qquad (5.7)\\
&\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} h_{11}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta)duds + \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} h_{21}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta)dW_u ds\\
&+\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} h_{12}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta)du dW_s + \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} h_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta)dW_u dW_s
\end{aligned}$$

where $h_{ij} = \mathcal{L}_i\mathcal{L}_j f,\ i,j=1,2$. In the two cases h_{11} and h_{12}, we can prove the result for $j=4$, which implies the result for $j=3$.

$$\begin{aligned}
&\frac{1}{\Delta_n^{4m}} E_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} Dh_{11}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)duds\right|^{2m}\right)\\
\le\ & \frac{1}{n\Delta_n^{4m}} \sum_{i=1}^n E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} Dh_{11}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)duds\right|^{2m}\right)\\
\le\ & \frac{1}{n\Delta_n^{2}} \sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} E_{\theta_0}\left(|Dh_{11}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)|^{2m}\right) duds\\
\le_C\ & |\theta_2-\theta_1|^{2m},
\end{aligned}$$

$$
\begin{aligned}
&\frac{1}{\Delta_n^{3m}} E_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} Dh_{21}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)dW_u ds\right|^{2m}\right)\\
&\quad \le \quad \frac{1}{\Delta_n^{3m} n}\sum_{i=1}^n E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} Dh_{21}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)dW_u ds\right|^{2m}\right)\\
&\quad \le \quad \frac{1}{\Delta_n^{m+1} n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{s} Dh_{21}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)dW_u\right|^{2m}\right) ds\\
&\quad \le_C \quad \frac{1}{\Delta_n^{m+1} n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n} E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{s} Dh_{21}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)^2 ds\right|^{m}\right) ds\\
&\quad \le \quad \frac{1}{\Delta_n^{2} n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} E_{\theta_0}\left(|Dh_{21}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)|^{2m} du ds\right)\\
&\quad \le_C \quad |\theta_2-\theta_1|^{2m},
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{\Delta_n^{4m}} E_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} Dh_{12}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)du dW_s\right|^{2m}\right)\\
&\quad \le_C \quad \frac{1}{\Delta_n^{4m}} E_{\theta_0}\left(\left|\frac{1}{n^2}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n}\left(\int_{t_{i-1}^n}^{s} Dh_{12}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)du\right)^2 ds\right|^{m}\right)\\
&\quad \le \quad \frac{1}{n^{m+1}\Delta_n^{4m}}\sum_{i=1}^n E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{t_i^n}\left(\int_{t_{i-1}^n}^{s} Dh_{12}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)du\right)^2 ds\right|^{m}\right)\\
&\quad \le \quad \frac{1}{n^{m+1}\Delta_n^{m+2}}\sum_{i=1}^n \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} E_{\theta_0}\left(|Dh_{12}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)|^{2m} du ds\right)\\
&\quad \le_C \quad \frac{1}{(n\Delta_n)^m}|\theta_2-\theta_1|^{2m},
\end{aligned}
$$

and

$$
\begin{aligned}
&\frac{1}{\Delta_n^{3m}} E_{\theta_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} Dh_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)dW_u dW_s\right|^{2m}\right)\\
\le_C \quad & \frac{1}{\Delta_n^{3m}} E_{\theta_0}\left(\left|\frac{1}{n^2}\sum_{i=1}^{n}\int_{t_{i-1}^n}^{t_i^n}\left(\int_{t_{i-1}^n}^{s} Dh_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)dW_u\right)^2 ds\right|^{m}\right)\\
\le \quad & \frac{1}{n^{m+1}\Delta_n^{3m}}\sum_{i=1}^{n} E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{t_i^n}\left(\int_{t_{i-1}^n}^{s} Dh_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)dW_u\right)^2 ds\right|^{m}\right)\\
\le \quad & \frac{1}{n^{m+1}\Delta_n^{2m+1}}\sum_{i=1}^{n}\int_{t_{i-1}^n}^{t_i^n} E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{s} Dh_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)dW_u\right|^{2m}\right) ds\\
\le_C \quad & \frac{1}{n^{m+1}\Delta_n^{2m+1}}\sum_{i=1}^{n}\int_{t_{i-1}^n}^{t_i^n} E_{\theta_0}\left(\left|\int_{t_{i-1}^n}^{s} Dh_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)^2 du\right|^{m}\right) ds\\
\le \quad & \frac{1}{n^{m+1}\Delta_n^{m+2}}\sum_{i=1}^{n}\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} E_{\theta_0}\left(|Dh_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta_2,\theta_1)|^{2m} du ds\right)\\
\le_C \quad & \frac{1}{(n\Delta_n)^m}|\theta_2-\theta_1|^{2m}.
\end{aligned}
$$

Finally, we prove (5.5) for $j = 4$. We have already taken care of two of the terms in (5.7), but the terms involving h_{21} and h_{22} require more work. Since $h_{2i} \in \mathcal{C}_0$, $i = 1, 2$, we find that

$$
\begin{aligned}
\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} h_{21}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta)dW_u ds \;=\; & \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s}\int_{t_{i-1}^n}^{u}\mathcal{L}_1 h_{21}(v-t_{i-1}^n, X_v, X_{t_{i-1}^n};\theta)dv dW_u ds\\
+\; & \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s}\int_{t_{i-1}^n}^{u}\mathcal{L}_2 h_{21}(v-t_{i-1}^n, X_v, X_{t_{i-1}^n};\theta)dW_v dW_u ds
\end{aligned}
$$

and

$$
\begin{aligned}
\int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s} h_{22}(u-t_{i-1}^n, X_u, X_{t_{i-1}^n};\theta)dW_u dW_s \;=\; & \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s}\int_{t_{i-1}^n}^{u}\mathcal{L}_1 h_{22}(v-t_{i-1}^n, X_v, X_{t_{i-1}^n};\theta)dv dW_u dW_s\\
+\; & \int_{t_{i-1}^n}^{t_i^n}\int_{t_{i-1}^n}^{s}\int_{t_{i-1}^n}^{u}\mathcal{L}_2 h_{22}(v-t_{i-1}^n, X_v, X_{t_{i-1}^n};\theta)dW_v dW_u dW_s.
\end{aligned}
$$

The result is now obtained by evaluating the triple integrals using the Burkholder-Davis-Gundy inequality and Jensen's inequality exactly as above. □

Lemma 5.5 *Under the Conditions 2.2 and 2.3*

$$
\frac{1}{n\Delta_n}\sum_{i=1}^{n} g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \xrightarrow{P_{\theta_0}} \gamma(\theta,\theta_0), \tag{5.8}
$$

$$\frac{1}{n\Delta_n}\sum_{i=1}^n \partial_{\theta^T} g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \xrightarrow{P_{\theta_0}} \tag{5.9}$$
$$\int_\ell^r [L_{\theta_0}(\partial_{\theta^T} g(0;\theta))(x,x) - L_\theta(\partial_{\theta^T} g(0;\theta))(x,x) - J_\theta(x)]\mu_{\theta_0}(x)dx,$$

and

$$\frac{1}{n\Delta_n}\sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^T \xrightarrow{P_{\theta_0}} V(\theta), \tag{5.10}$$

uniformly for θ in a compact set. The function γ given by (3.1) is a continuous function of θ. For a martingale estimating function or more generally if $n\Delta_n^{2\kappa-1} \to 0$,

$$\frac{1}{\sqrt{n\Delta_n}}\sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta_0) \xrightarrow{\mathcal{D}} N_2(0, V_0). \tag{5.11}$$

Proof. By (2.1), (2.9), (2.11) and Lemma 5.1,

$$\begin{aligned}
&E_{\theta_0}\left(g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \mid X_{t_{i-1}^n}\right)\\
&= \Delta_n\left[g^{(1)}(X_{t_{i-1}^n}, X_{t_{i-1}^n};\theta) + L_{\theta_0}(g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n})\right] + \Delta_n^2 R(\Delta_n, X_{t_{i-1}^n}, \theta)\\
&= \Delta_n\left[L_{\theta_0}(g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n}) - L_\theta(g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n})\right] + \Delta_n^2 R(\Delta_n, X_{t_{i-1}^n}, \theta).
\end{aligned}$$

The last equality follows from (2.12). Thus

$$\begin{aligned}
&\frac{1}{n\Delta_n}\sum_{i=1}^n E_{\theta_0}\left(g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \mid X_{t_{i-1}^n}\right)\\
&= \frac{1}{n}\sum_{i=1}^n\left[L_{\theta_0}(g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n}) - L_\theta(g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n})\right] + \Delta_n\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta)\\
&\xrightarrow{P_{\theta_0}} \gamma(\theta, \theta_0)
\end{aligned}$$

by Lemma 5.2. Moreover, $E_{\theta_0}\left(g_j(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^2 \mid X_{t_{i-1}^n}\right) = \Delta_n R(\Delta_n, X_{t_{i-1}^n}, \theta)$, so

$$\frac{1}{(n\Delta_n)^2}\sum_{i=1}^n E_{\theta_0}\left(g_j(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^2 \mid X_{t_{i-1}^n}\right) = \frac{1}{n\Delta_n}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta) \xrightarrow{P_{\theta_0}} 0.$$

Therefore pointwise convergence in (5.8) follows from Lemma 5.3. In order to prove that the convergence is uniform for θ in a compact set K, we show that the sequence $\zeta_n(\cdot) = \frac{1}{n\Delta_n}\sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}, \cdot)$ converges weakly to the limit $\gamma(\cdot, \theta_0)$ in the space, $C(K)$, of continuous functions on K with the supremum norm. Since the limit is non-random, this implies uniform convergence in probability for $\theta \in K$. That $\gamma(\cdot, \theta_0)$ is continuous follows from the dominated convergence theorem because of the imposed uniform polynomial growth assumptions. Since pointwise convergence has been established, weak convergence follows because the family of distributions of $\zeta_n(\cdot)$ is tight. The tightness follows from Lemma 5.4

with $f = g_i$, $j = 2$ and $m = 2$. That (5.5) and pointwise convergence implies tightness follows from Corollary 14.9 in Kallenberg (1997), which is a generalization of Theorem 12.3 in Billingsley (1968) (see also Lemma 3.1 in Yoshida (1990) and Theorem 20 in Appendix I of Ibragimov and Has'minskii (1981)).

In a similar way it follows from (2.1), (2.9), (2.11), (2.12) and Lemma 5.1 that

$$\begin{aligned} & E_{\theta_0}\left(\partial_{\theta^T} g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) \,|\, X_{t_{i-1}^n}\right) \qquad (5.12)\\ = \;& \Delta_n \left[\partial_{\theta^T} g^{(1)}(X_{t_{i-1}^n}, X_{t_{i-1}^n}; \theta) + L_{\theta_0}(\partial_{\theta^T} g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n})\right] + \Delta_n^2 R(\Delta_n, X_{t_{i-1}^n}, \theta) \\ = \;& \Delta_n \left[L_{\theta_0}(\partial_{\theta^T} g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n}) - L_{\theta}(\partial_{\theta^T} g(0;\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n}) - J_\theta(X_{t_{i-1}^n})\right] \\ & \qquad + \Delta_n^2 R(\Delta_n, X_{t_{i-1}^n}, \theta), \end{aligned}$$

and from (2.1),(2.9), (2.11), and Lemma 5.1 that

$$\begin{aligned} & E_{\theta_0}\left(g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta)^T \,|\, X_{t_{i-1}^n}\right) \\ = \;& \Delta_n v(X_{t_{i-1}^n}, \beta_0) \partial_y g(0, X_{t_{i-1}^n}, X_{t_{i-1}^n}; \theta) \partial_y g(0, X_{t_{i-1}^n}, X_{t_{i-1}^n}; \theta)^T + \Delta_n^2 R(\Delta_n, X_{t_{i-1}^n}, \theta). \end{aligned}$$

Since by (2.1),(2.9), (2.11), and Lemma 5.1

$$E_{\theta_0}\left([\partial_\theta g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta)]^2 \,|\, X_{t_{i-1}^n}\right) = \Delta_n R(\Delta_n, X_{t_{i-1}^n}, \theta)$$

and

$$E_{\theta_0}\left([g_j(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) g_k(\Delta, X_{t_i^n}, X_{t_{i-1}^n}; \theta)]^2 \,|\, X_{t_{i-1}^n}\right) = \Delta_n R(\Delta_n, X_{t_{i-1}^n}, \theta), \qquad (5.13)$$

we can, as above, use Lemma 5.2 and Lemma 5.3 to prove (5.9) and (5.10). As above, uniform convergence for θ in a compact set K follows by using Lemma 5.4 with $f = \partial_{\theta_j} g_k$ and $f = g_j g_k$ to prove the tightness of (5.4) (with $j = 2$) in $C(K)$.

Finally, (5.11) follows from the central limit theorem for square integrable martingale arrays under conditions which, in the martingale case, we have already verified in the proof of (5.10), see e.g. Corollary 3.1 in Hall and Heyde (1980) with the conditional Lindeberg condition replaced by the stronger conditional Liapounov condition that follows from (5.13) and Lemma 5.2, e.g.

$$\frac{1}{(n\Delta_n)^2} \sum_{i=1}^n E_{\theta_0}\left(g_j(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta_0)^4 \,|\, X_{t_{i-1}^n}\right) = \frac{1}{n\Delta_n} \frac{1}{n} \sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta_0) \xrightarrow{P_{\theta_0}} 0.$$

The nestedness condition in Hall and Heyde's Corollary 3.1 is not needed here because the limit of the quadratic variation is non-random.

In the case of non-martingale estimating functions, we consider the martingale $\sum_{i=1}^n \tilde{g}(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta_0)$, where $\tilde{g} = g - E_{\theta_0}\left(g |\, X_{t_{i-1}^n}\right)$. This martingale satisfies the conditions of the central limit theorem, which follows from the expansions of conditional expectations given above and $E_{\theta_0}\left(g_j(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta_0)^3 \,|\, X_{t_{i-1}^n}\right) = \Delta_n R(\Delta_n, X_{t_{i-1}^n}, \theta_0)$. Now, (5.11) follows because by (2.7)

$$\frac{1}{\sqrt{n\Delta_n}}\sum_{i=1}^n E_{\theta_0}\left(g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta_0)\,|\,X_{t_{i-1}^n}\right) = \sqrt{n}\Delta_n^{\kappa-1/2}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n},\theta_0) \xrightarrow{P_{\theta_0}} 0. \tag{5.14}$$

□

Proof of Theorem 3.1. By Lemma 5.5, the estimating function

$$G_n(\theta) = \frac{1}{n\Delta_n}\sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \tag{5.15}$$

satisfies the conditions that $G_n(\theta_0) \overset{P_{\theta_0}}{\to} 0$, $\partial_\theta G_n(\theta) \overset{P_{\theta_0}}{\to} U(\theta)$ uniformly for θ in a compact set, and that $U(\theta_0) = -S$ is invertible. Here $U(\theta)$ denotes the right hand side of (5.9). This implies the eventual existence and the consistency of $\hat{\theta}_n$ as well as the eventual uniqueness of consistent G_n-estimators; see Theorems 2.5 and 2.6 in Jacod and Sørensen (2018). Now consider any G_n-estimator $\tilde{\theta}_n$ for which $P_{\theta_0}(\tilde{\theta}_n \in K) \to 1$ as $n \to \infty$, where K is a compact subset of Θ with $\theta_0 \in \operatorname{int} K$. By Theorem 2.7 in Jacod and Sørensen (2018) the facts that $\gamma(\theta,\theta_0)$ (the limit of $G_n(\theta)$) satisfies that $\gamma(\theta,\theta_0) \neq 0$ for $\theta \neq \theta_0$ and is continuous in θ imply that $P_{\theta_0}(\hat{\theta}_n \neq \tilde{\theta}_n) \to 0$ as $n \to \infty$. The asymptotic normality follows by standard arguments. Finally, (3.6) and (3.7) follow from (5.9) and (5.10) because the convergence is uniform for θ in compact sets. □

The next lemma is needed in the proof of Theorem 3.2

Lemma 5.6 *Under the Conditions 1.1, 2.2, and 2.3*

$$D_n\sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^T D_n \xrightarrow{P_{\theta_0}} \begin{pmatrix} W_1(\theta) & 0 \\ 0 & W_2(\theta)\end{pmatrix} \tag{5.16}$$

uniformly for θ in a compact set, where D_n is given by (3.11).

For a martingale estimating function or if more generally $n\Delta^{2(\kappa-1)} \to 0$,

$$\begin{pmatrix} \frac{1}{\sqrt{n\Delta_n}}\sum_{i=1}^n g_1(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta_0) \\ \frac{1}{\Delta_n\sqrt{n}}\sum_{i=1}^n g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta_0)\end{pmatrix} \xrightarrow{\mathcal{D}} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} W_1(\theta_0) & 0 \\ 0 & W_2(\theta_0)\end{pmatrix}\right). \tag{5.17}$$

If, in addition, condition (3.8) holds, then

$$\frac{1}{n\Delta_n^{3/2}}\sum_{i=1}^n \partial_\alpha g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) \xrightarrow{P_{\theta_0}} 0 \tag{5.18}$$

uniformly for θ in a compact set.

Proof. By (2.1), (2.9), (2.11), (1.3) and Lemma 5.1,

$$\frac{1}{n\Delta_n^{3/2}}\sum_{i=1}^n E_{\theta_0}\left(g_1(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta) g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)\,|\,X_{t_{i-1}^n}\right)$$

$$= \Delta_n^{1/2}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n},\theta) \xrightarrow{P_{\theta_0}} 0$$

and

$$\frac{1}{n^2\Delta_n^3}\sum_{i=1}^n E_{\theta_0}\left([g_1(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)]^2 \,|\, X_{t_{i-1}^n}\right) \tag{5.19}$$
$$= \frac{1}{n\Delta_n}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta) \xrightarrow{P_{\theta_0}} 0,$$

so the pointwise convergence of the two off-diagonal entries in (5.16) follows from Lemma 5.3. Similarly to the proof of Lemma 5.5, uniform convergence for θ in a compact set K follows by using Lemma 5.4 with $f = g_1 g_2$ to prove the tightness of (5.4) (with $j = 3$) in $C(K)$.

The convergence of $(n\Delta_n)^{-1}\sum_{i=1}^n g_1(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^2$ was taken care of in Lemma 5.5. By (2.1), (2.9), (2.11), (2.12), (1.3) and Lemma 5.1, we see that

$$\begin{aligned}
&E_{\theta_0}\left(g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^2 \,|\, X_{t_{i-1}^n}\right)\\
&= \Delta_n^2\Big[\tfrac{1}{2}L_{\theta_0}^2(g_2(0;\theta)^2)(X_{t_{i-1}^n}, X_{t_{i-1}^n}) + 2L_{\theta_0}(g_2(0;\theta)g_2^{(1)}(\theta))(X_{t_{i-1}^n}, X_{t_{i-1}^n})\\
&\qquad + g_2^{(1)}(X_{t_{i-1}^n}, X_{t_{i-1}^n};\theta)^2\Big] + \Delta_n^3 R(\Delta_n, X_{t_{i-1}^n}, \theta)\\
&= \tfrac{1}{2}\Delta_n^2\left[v(X_{t_{i-1}^n};\beta_0)^2 + \tfrac{1}{2}(v(X_{t_{i-1}^n};\beta_0) - v(X_{t_{i-1}^n};\beta))^2\right](\partial_y^2 g_2(0, X_{t_{i-1}^n}, X_{t_{i-1}^n};\theta))^2\\
&\qquad + \Delta_n^3 R(\Delta_n, X_{t_{i-1}^n}, \theta),
\end{aligned}$$

Thus

$$\begin{aligned}
&\frac{1}{n\Delta_n^2}\sum_{i=1}^n E_{\theta_0}\left(g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^2 \,|\, X_{t_{i-1}^n}\right)\\
&= \frac{1}{n}\sum_{i=1}^n \tfrac{1}{2}\left[v(X_{t_{i-1}^n};\beta_0) + \tfrac{1}{2}(v(X_{t_{i-1}^n};\beta_0) - v(X_{t_{i-1}^n};\beta))^2\right](\partial_y^2 g_2(0, X_{t_{i-1}^n}, X_{t_{i-1}^n};\theta))^2\\
&\qquad + \Delta_n \frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta)\\
&\xrightarrow{P_{\theta_0}} W_2(\theta)
\end{aligned}$$

by Lemma 5.2. We conclude that $(n\Delta_n^2)^{-1}\sum_{i=1}^n g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}\theta)^2$ converges to $W_2(\theta)$ by Lemma 5.3 because

$$\frac{1}{n^2\Delta_n^4}\sum_{i=1}^n E_{\theta_0}\left(g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n};\theta)^4 \,|\, X_{t_{i-1}^n}\right) = \frac{1}{n\Delta_n}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta) \xrightarrow{P_{\theta_0}} 0. \tag{5.20}$$

This follows from (2.1), (2.9), (2.11), (1.3), and Lemmas 5.1 and 5.2. Uniform convergence for θ in a compact set K follows by using Lemma 5.4 with $f = g_2^2$ to prove the tightness of (5.4) (with $j = 4$) in $C(K)$.

As in the proof of Lemma 5.5, (5.17) follows from the central limit theorem for square integrable martingale arrays (Corollary 3.1 in Hall and Heyde (1980)) under conditions which, in the martingale case, we have already verified in the proof of (5.16). In particular, the conditional Liapounov condition follows from (5.13), (5.20) and (5.19).

In the case of non-martingale estimating functions, consider the martingale $\sum_{i=1}^n \tilde{g}(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta_0)$, where $\tilde{g} = g - E_{\theta_0}\left(g | X_{t_{i-1}^n}\right)$, which satisfies the conditions of the central limit theorem. This follows from the expansions of conditional expectations given above and $E_{\theta_0}\left(g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta_0)^3 \,|\, X_{t_{i-1}^n}\right) = \Delta_n^2 R(\Delta_n, X_{t_{i-1}^n}, \theta_0)$. Now, (5.17) follows because g_1 satisfies (5.14), and because by (2.7)

$$\frac{1}{\Delta_n\sqrt{n}}\sum_{i=1}^n E_{\theta_0}\left(g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta_0) \,|\, X_{t_{i-1}^n}\right) = \sqrt{n}\Delta_n^{\kappa-1}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta_0) \xrightarrow{P_{\theta_0}} 0.$$

Finally, to prove (5.18) note that (5.12), (1.3) and (3.8) imply that

$$E_{\theta_0}\left(\partial_\alpha g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) \,|\, X_{t_{i-1}^n}\right) = \Delta^2 R(\Delta_n, X_{t_{i-1}^n}, \theta),$$

and that it follows from (2.1), (2.9), (2.11), (2.12), (1.3), (3.8) and Lemma 5.1 that

$$E_{\theta_0}\left([\partial_\alpha g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta)]^2 \,|\, X_{t_{i-1}^n}\right) = \Delta_n^3 R(\Delta_n, X_{t_{i-1}^n}, \theta).$$

Therefore by Lemma 5.2

$$\frac{1}{n\Delta_n^{3/2}}\sum_{i=1}^n E_{\theta_0}\left(\partial_\alpha g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta) \,|\, X_{t_{i-1}^n}\right) = \sqrt{\Delta_n}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}; \theta) \xrightarrow{P_{\theta_0}} 0.$$

and

$$\frac{1}{n^2\Delta_n^3}\sum_{i=1}^n E_{\theta_0}\left([\partial_\alpha g_2(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta)]^2 \,|\, X_{t_{i-1}^n}\right) = \frac{1}{n}\frac{1}{n}\sum_{i=1}^n R(\Delta_n, X_{t_{i-1}^n}, \theta) \xrightarrow{P_{\theta_0}} 0,$$

so that (5.18) follows from Lemma 5.3. Uniform convergence for θ in a compact set K follows by using Lemma 5.4 with $f = \partial_\alpha g_2$ to conclude tightness of (5.4) (with $j = 3$) in $C(K)$. To see that $\partial_\alpha g_2$ satisfies the conditions of the lemma, we use (2.12) and (3.8) to conclude that $\partial_\Delta \partial_\alpha g_2(0, x, x; \theta) = \partial_\alpha g_2^{(1)}(x, x; \theta) = -\partial_\alpha L_\theta(g_2(0; \theta))(x, x) = 0$.

□

Proof of Theorem 3.2. The results on eventual existence, uniqueness and consistence of $\hat{\theta}_n$ follow from Theorem 3.1: Because (1.3) implies $S_{21} = 0$, the assumptions that $S_{11} \neq 0$ and $S_{22} \neq 0$ ensure that S is invertible, and similarly, under Condition 1.1 the identifiability condition imposed in Theorem 3.2 ensures that $\gamma(\theta, \theta_0) \neq 0$ for $\theta \neq \theta_0$, where γ is the limit of $G_n(\theta)$ given by (3.1).

To prove the asymptotic normality (3.9) of the estimator $\hat{\theta}_n$ we consider

$$\tilde{G}_n(\theta) = D_n \sum_{i=1}^n g(\Delta_n, X_{t_i^n}, X_{t_{i-1}^n}; \theta),$$

where D_n is given by (3.11). On the set $\{\tilde{G}_n(\hat{\theta}_n) = 0\}$ (the probability of which goes to one)

$$-\partial_{\theta^T}\tilde{G}_n(\theta_n^{(1)}, \theta_n^{(2)}) A_n^{-1} A_n(\hat{\theta}_n - \theta_0) = \tilde{G}_n(\theta_0),$$

where

$$A_n = \begin{pmatrix} \sqrt{\Delta_n n} & 0 \\ 0 & \sqrt{n} \end{pmatrix},$$

$\partial_{\theta^T}\tilde{G}_n(\theta_n^{(1)},\theta_n^{(2)})$ is the 2×2-matrix whose jkth entry is $\partial_{\theta_k}\tilde{G}_n(\theta_n^{(j)})_j$, and $\theta_n^{(j)}$ is a random convex combination of $\hat{\theta}_n$ and θ_0. Since by (5.9) and (5.18)

$$-\partial_{\theta^T}\tilde{G}_n(\theta_n^{(1)},\theta_n^{(2)})A_n^{-1} \xrightarrow{P_{\theta_0}} \begin{pmatrix} S_{11} & 0 \\ 0 & S_{22} \end{pmatrix},$$

(3.9) follows from (5.17).

Finally, (3.10) follows from (5.16) because the convergence is uniform for θ in compact sets.

□

Proof of Theorem 4.1. That the conclusions of Theorem 3.2 holds is obvious because (1.5) implies (3.8). The efficiency follows from Theorem 4.1 in Gobet (2002), where it is proved that the diffusion model (1.1) is locally asymptotically normal with Fisher information matrix $\mathcal{I}(\theta_0) = \Sigma(\theta_0)^{-1}$, with $\Sigma(\theta_0)$ given by (4.1). Under Condition 1.2, $S_{11} = W_1(\theta_0) = \mathcal{I}_{11}(\theta_0)$ and $S_{22} = W_2(\theta_0) = \mathcal{I}_{22}(\theta_0)$, so it follows from (3.9) that the asymptotic covariance matrix of $\hat{\theta}_n$ equals the inverse of the Fisher information matrix. The estimators of the asymptotic variances follow from (3.10).

□

Proof of Theorem 4.4. By (2.9) $\pi_\theta^{1,\Delta} f(x;\theta) = f(x;\theta) + \Delta L_\theta f(x;\theta) + O(\Delta^2)$, so after another application of (2.9), we see that $h(\Delta,y,x;\theta) = f(y;\theta) - \pi_\theta^{\kappa,\Delta} f(x;\theta)$, $\kappa \in \mathbb{N}$, satisfies

$$\begin{aligned} E_\theta\left(h(\Delta,X_\Delta,x;\theta)h(\Delta,X_\Delta,x;\theta)^T \mid X_0 = x\right) &= \Delta L_\theta(h(0;\theta)h(0;\theta)^T)(x,x) \\ &\quad + \Delta^2\left(\tfrac{1}{2}L_\theta^2(h(0;\theta)h(0;\theta)^T)(x,x) - L_\theta f(x;\theta)L_\theta f^T(x;\theta)\right) + O(\Delta^3) \\ &= \Delta v(x;\beta)\partial_x f(x;\theta)\partial_x f(x;\theta)^T + \Delta^2 K(x) + O(\Delta^3), \end{aligned}$$

where

$$\begin{aligned} K(x) &= q_1(x;\theta)\partial_x f(x;\theta)\partial_x f(x;\theta)^T + q_2(x;\theta)\left(\partial_x^2 f(x;\theta)\partial_x f(x;\theta)^T + \partial_x f(x;\theta)\partial_x^2 f(x;\theta)^T\right) \\ &\quad + v(x;\beta)^2\left(\partial_x^2 f(x;\theta)\partial_x^2 f(x;\theta)^T + \tfrac{1}{2}(\partial_x^3 f(x;\theta)\partial_x f(x;\theta)^T + \partial_x f(x;\theta)\partial_x^3 f(x;\theta)^T)\right) \end{aligned}$$

with

$$\begin{aligned} q_1(x;\theta) &= \tfrac{1}{2}[b(x;\alpha)(2+\partial_x v(x;\beta)) - 2b(x;\alpha) + \tfrac{1}{2}v(x;\beta)(4\partial_x b(x;\alpha) + \partial_x^2 v(x;\beta))] \\ q_2(x,\theta) &= \tfrac{3}{4}v(x;\beta)(1 + \tfrac{1}{3}b(x;\alpha) + \partial_x v(x;\beta)). \end{aligned}$$

Since

$$\begin{aligned} \partial_\alpha L_\theta f(x;\theta) - L_\theta\partial_\alpha f(x;\theta) &= \partial_\alpha b(x;\alpha)\partial_x f(x;\theta) \\ \partial_\beta L_\theta f(x;\theta) - L_\theta\partial_\beta f(x;\theta) &= \tfrac{1}{2}\partial_\beta v(x;\beta)\partial_x^2 f(x;\theta), \end{aligned}$$

it also follows from (2.9) that

$$\partial_{\theta^T}\pi_\theta^\Delta f(x) - \pi_\theta^\Delta\partial_{\theta^T} f(x) = \Delta F(x)\begin{pmatrix} \partial_\alpha b(x;\alpha) & 0 \\ 0 & \tfrac{1}{2}\partial_\beta v(x;\beta) \end{pmatrix} + O(\Delta^2),$$

where $F(x)$ denotes the $N \times 2$-matrix $F(x) = (\partial_x f(x), \partial_x^2 f(x))$.

If $A^*(x,\Delta;\theta)$ satisfies (4.8), then the $2 \times N$-matrix

$$B(x,\Delta;\theta) = \begin{pmatrix} 1 & 0 \\ 0 & 2\Delta \end{pmatrix} A^*(x,\Delta;\theta).$$

satisfies that

$$B(x,\Delta;\theta)\left[v(x;\beta)\partial_x f(x;\theta)\partial_x f(x;\theta)^T + \Delta K(x;\theta) + O(\Delta^2)\right] \tag{5.21}$$

$$= \begin{pmatrix} \partial_\alpha b(x;\alpha) & 0 \\ 0 & \Delta\partial_\beta v(x;\beta) \end{pmatrix} F(x)^T + \begin{pmatrix} O(\Delta) \\ O(\Delta^2) \end{pmatrix}.$$

Let $B(x,\Delta;\theta)_i$ denote the ith row of $B(x,\Delta;\theta)$ $(i = 1,2)$. Then it follows by letting Δ tend to zero that

$$v(x;\beta)B(x,0;\theta)_2\partial_x f(x;\theta)\partial_x f(x;\theta)^T = 0. \tag{5.22}$$

The condition that $M(x)$ is invertible implies that we can find a coordinate of $\partial_x f(x;\theta)$ which is not equal to zero, so we conclude that

$$\partial_y g_2^*(0,x,x;\theta) = B(x,0;\theta)_2\partial_x f(x;\theta) = 0.$$

Similarly we find that

$$[v(x;\beta)B(x,0;\theta)_1\partial_x f(x;\theta) - \partial_\alpha b(x;\alpha)]\partial_x f(x;\theta)^T = 0,$$

which implies

$$\partial_y g_1^*(0,x,x;\theta) = B(x,0;\theta)_1\partial_x f(x;\theta) = \partial_\alpha b(x;\alpha)/v(x;\beta).$$

Finally, (5.21) and (5.22) imply that $B(x,\Delta;\theta)_2\left(\Delta K(x;\theta) + O(\Delta^2)\right) = \Delta\partial_\beta v(x;\beta)\partial_x^2 f(x;\theta)^T + O(\Delta^2)$, so that

$$B(x,0;\theta)_2 K(x;\theta) = \partial_\beta v(x;\beta)\partial_x^2 f(x;\theta)^T.$$

Since we have shown that $B(x,0;\theta)_2\partial_x f(x;\theta) = 0$, this expression can be rewritten as

$$c_1(x;\theta)\partial_x f(x;\theta) = c_2(x;\theta)\partial_x^2 f(x;\theta)$$

where

$$\begin{aligned} c_1(x;\theta) &= q_2(x;\theta)B(x,0;\theta)_2\partial_x^2 f(x;\theta) + \tfrac{1}{2}v(x;\beta)^2 B(x,0;\theta)_2\partial_x^3 f(x;\theta) \\ c_2(x;\theta) &= \partial_\beta v(x;\beta) - v(x;\beta)^2 B(x,0;\theta)_2\partial_x^2 f(x;\theta). \end{aligned}$$

If $c_2(x;\theta) \neq 0$, then $\partial_x^2 f(x) = c_1(x;\theta)/c_2(x;\theta)\partial_x f(x;\theta)$, which implies that $\det(M(x)) = 0$. This contradicts the assumption that $M(x)$ is invertible, so we conclude that $\partial_\beta v(x;\beta) - v(x;\beta)^2 B(x,0;\theta)_2\partial_x^2 f(x;\theta) = 0$ or equivalently

$$\partial_y^2 g_2^*(0,x,x;\theta) = B(x,0;\theta)_2\partial_x^2 f(x;\theta) = \partial_\beta v(x;\beta)/v(x;\beta)^2.$$

That the results hold for $\tilde{B}$ is obvious. Note that it also follows that neither $B(x,0;\theta)_1$ nor $B(x,0;\theta)_2$ is the zero vector, so there exist x and y such that $g_i^*(0,y,x;\theta) \neq 0$, $i = 1,2$. □

Acknowledgements

I am grateful to Arnaud Gloter, Nina Munkholt Jakobsen, Mathieu Kessler, Masayuki Uchida and Nakahiro Yoshida for helpful discussions.

References

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, **70**, 223–262.

Aït-Sahalia, Y. and Mykland, P. (2003). The effects of random and discrete sampling when estimating continuous-time diffusions. *Econometrica*, **71**, 483–549.

Arnst, M., Louppe, G., Hulle, R. V., L. Gillet, F. B., and Denoöel, V. (2022). A hybrid stochastic model and its bayesian identification for infectious disease screening in a university campus with application to massive covid-19 screening at the university of liège. *Mathematical Biosciences*, **347**, 108805.

Bachar, M., Batzel, J., and Ditlevsen, S. (2013). *Stochastic Biomathematical Models*. Springer, Heidelberg.

Barndorff-Nielsen, O. E. and Sørensen, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic processes. *International Statistical Review*, **62**, 133–165.

Beskos, A., Papaspiliopoulos, O., and Roberts, G. O. (2009). Monte carlo maximum likelihood estimation for discretely observed diffusion processes. *Ann. Statist.*, **37**, 223–245.

Bibbona, E., Lansky, P., and Sirovich, R. (2010). Estimating input parameters from intracellular recordings in the Feller neuronal model. *Physical Review E*, **81**, 031916.

Bibby, B. M. and Sørensen, M. (1995). Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, **1**, 17–39.

Bibby, B. M. and Sørensen, M. (1996). On estimation for discretely observed diffusions: A review. *Theory of Stochastic Processes*, **2**, 49–56.

Bibby, B. M., Jacobsen, M., and Sørensen, M. (2010). Estimating functions for discretely sampled diffusion-type models. In Y. Aït-Sahalia and L. P. Hansen, editors, *Handbook of Financial Econometrics*, pages 203–268. North Holland, Oxford.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Bladt, M. and Sørensen, M. (2014). Simple simulation of diffusion bridges with application to likelihood inference for diffusions. *Bernoulli*, **20**, 645 – 675.

Bladt, M., Mider, M., and Sørensen, M. (2021). Corrigendum to "Simple simulation of diffusion bridges with application to likelihood inference for diffusions". *Bernoulli*, **27**, 218 – 220.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton University Press, Princeton.

Chan, K. C., Karolyi, G. A., Longstaff, F. A., and Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance*, **47**, 1209–1227.

Christensen, B. J. and Sørensen, M. (2008). Optimal inference in dynamic models with conditional moment restrictions. Preprint, Creates.

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

Dipple, S., Choudhary, A., Flamino, J., Szymanski, B. K., and Korniss, G. (2020). Using correlated stochastic differential equations to forecast cryptocurrency rates and social media activities. *Applied Network Science*, **5**, 17.

Ditlevsen, P. and Ditlevsen, S. (2023). Warning of a forthcoming collapse of the Atlantic meridional overturning circulation. *Nature Communications*, **14**, 4254.

Dorogovcev, A. J. (1976). The consistency of an estimate of a parameter of a stochastic differential equation. *Theor. Probability and Math. Statist.*, **10**, 73–82.

Durham, G. B. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *J. Business & Econom. Statist.*, **20**, 297–338.

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

Forman, J. L. and Sørensen, M. (2008). The Pearson diffusions: a class of statistically tractable diffusion processes. *Scandinavian Journal of Statistics*, **35**, 438–465.

Genon-Catalot, V. and Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. I. H. Poincaré PR*, **29**, 119–151.

Gihman, I. I. and Skorohod, A. V. (1972). *Stochastic Differential Equations*. Springer-Verlag, Berlin.

Gloter, A. and Sørensen, M. (2009). Estimation for stochastic differential equations with a small diffusion coefficient. *Stoch. Proc. Appl.*, **119**, 679–699.

Gobet, E. (2002). Lan property for ergodic diffusions with discrete observations. *Ann. I. H. Poincaré PR*, **38**, 711–737.

Godambe, V. P. and Heyde, C. C. (1987). Quasi likelihood and optimal estimation. *International Statistical Review*, **55**, 231–244.

Golden, M., García-Portugués, E., Sørensen, M., Mardia, K. V., Hamelryck, T., and Hein, J. (2017). A generative angular model of protein structure evolution. *Mol. Biol. Evol.*, **43**, 2085 – 2100.

Golightly, A. and Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle mcmc. *Interface Focus*, **1**, 807–820.

Guy, R., Larédo, C., and Vergu, E. (2015). Approximation of epidemic models by diffusion processes and their statistical inference. *Mathematical Biology*, **70**, 621 – 646.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and its Applications*. Academic Press, New York.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics*, **30**, 203–238.

Hansen, L. P. (1993). Semiparametric efficiency bounds for linear time-series models. In P. C. B. Phillips, editor, *Models, Methods and Applications of Econometrics: Esseys in Honor of A. R. Bergstrom*, pages 253–271. Cambridge, MA: Blackwell.

Hansen, L. P. (2001). Method of moments. In *International Encyclopedia of the Social and Behavior Sciences*. New York: Elsevier.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Application*. Springer-Verlag, New York.

Iacus, S. M. and Yoshida, N. (2018). *Simulation and Inference for Stochastic Processes with YUIMA*. Springer, Cham.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation - Asymptotic Theory*. Springer-Verlag, New York.

Jacobsen, M. (2001). Discretely observed diffusions; classes of estimating functions and small δ-optimality. *Scand. J. Statist.*, **28**, 123–150.

Jacobsen, M. (2002). Optimality and small δ-optimality of martingale estimating functions. *Bernoulli*, **8**, 643–668.

Jacod, J. and Sørensen, M. (2018). A review of asymptotic theory of estimating functions. *Statistical Inference for Stochastic processes*, **21**, 415–434.

Jakobsen, N. M. and Sørensen, M. (2017). Efficient estimation for diffusions sampled at high frequency over a fixed time interval. *Bernoulli*, **23**, 1874–1910.

Jensen, A. C., Ditlevsen, S., Kessler, M., and Papaspiliopoulos, O. (2012). Markov chain monte carlo approach to parameter estimation in the FitzHugh-Nagumo model. *Physical Review*, **E86**, 041114.

Jørgensen, E. S. and Sørensen, M. (2021). Prediction-based estimation for diffusion models with high-frequency data. *Japanese Journal of Statistics and Data Science*, **4**, 483–511.

Kallenberg, O. (1997). *Foundations of Modern Probability*. Springer-Verlag, New York.

Kelly, L., Platen, E., and Sørensen, M. (2004). Estimation for discretely observed diffusions using transform functions. *J. Appl. Prob.*, **41**, 99–118.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.*, **24**, 211–229.

Kessler, M. and Sørensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, **5**, 299–314.

Kitagawa, H. and Uchida, M. (2014). Adaptive test statistics for ergodic diffusion processes sampled at discrete times. *Journal of Statistical Planning and Inference*, **150**, 84–110.

Kutoyants, Y. (2004). *Statistical Inference for Ergodic Diffusion Processes*. Springer, New York.

Larsen, K. S. and Sørensen, M. (2007). A diffusion model for exchange rates in a target zone. *Mathematical Finance*, **17**, 285–306.

Michelot, T., Gloaguen, P., Blackwell, P. G., and Étienne, M.-P. (2019). The Langevin diffusion as a continuous-time model of animal movement and habitat selection. *Methods in Ecology and Evolution*, **10**, 1894 – 1907.

Overbeck, L. and Rydén, T. (1997). Estimation in the cox-ingersoll-ross model. *Econometric Theory*, **13**, 430–461.

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, **22**, 55–71.

Picchini, U., Ditlevsen, S., and Gaetano, A. D. (2008). Maximum likelihood estimation of a time-inhomogeneous stochastic differential model of glucose dynamics. *Mathematical Medicine and Biology*, **25**, 141–155.

Prakasa Rao, B. L. S. (1988). Statistical inference from sampled data for stochastic processes. *Contemporary Mathematics*, **80**, 249–284.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using metropolis-hastings algorithms. *Biometrika*, **88**, 603–621.

Sørensen, M. and Uchida, M. (2003). Small-diffusion asymptotics for discretely sampled stochastic differential equations. *Bernoulli*, **9**, 1051–1069.

Sørensen, M. (2012). Estimating functions for diffusion-type processes. In M. Kessler, A. Lindner, and M. Sørensen, editors, *Statistical Methods for Stochastic Differential Equations*, pages 1–107. Chapman & Hall/CRC Press.

Uchida, M. (2010). Contrast-based information criterion for ergodic diffusion processes from discrete observations. *Ann. Inst. Stat. Math.*, **62**, 161–187.

Uchida, M. and Yoshida, N. (2012). Adaptive estimation of an ergodic diffusion process based on sampled data. *Stochastic Processes and Their Applications*, **122**, 2885–2924.

Uchida, M. and Yoshida, N. (2013). Quasi likelihood analysis of volatility and nondegeneracy of statistical random fields. *Stochastic Processes and Their Applications*, **123**, 2851–2876.

van der Meulen, F. and Schauer, M. (2017). Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronic Journal of Statistics*, **11**, 2358 – 2396.

Yoshida, N. (1990). Asymptotic behaviour of m-estimator and related random field for diffusion process. *Ann. Inst. Statist. Math*, **42**, 221–251.

Yoshida, N. (1992). Estimation for diffusion processes from discrete observations. *Journal of Multivariate Analysis*, **41**, 220–242.