ASYMPTOTIC PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATION: PARAMETERISED DIFFUSION IN A MANIFOLD

S. SAID,* *The University of Melbourne*

J. H. MANTON,** *The University of Melbourne*

Abstract

This paper studies maximum likelihood estimation for a parameterised elliptic diffusion in a manifold. The focus is on asymptotic properties of maximum likelihood estimates obtained from continuous time observation. These are well known when the underlying manifold is a Euclidean space. However, no systematic study exists in the case of a general manifold. The starting point is to write down the likelihood function and equation. This is achieved using the tools of stochastic differential geometry. Consistency, asymptotic normality and asymptotic optimality of maximum likelihood estimates are then proved, under regularity assumptions. Numerical computation of maximum likelihood estimates is briefly discussed.

Keywords: maximum likelihood, elliptic diffusion, differentiable manifold, Fisher information, asymptotic normality

2010 Mathematics Subject Classification: Primary 93E10
Secondary 58J65

1. Introduction

Diffusions in manifolds, especially in classical matrix Lie groups and symmetric spaces, are natural models for many engineering problems. Such problems range from the control of robots or vehicles to the computational dynamics of large molecules [1,2].

Historically, one of the earliest studies of diffusions in matrix manifolds was concerned with rotation Brownian motion as a model for molecular dynamics — See Perrin's 1928 paper [3]. After the advent of stochastic calculus, there was renewed interest in diffusions in manifolds, pioneered by Yosida [4] and Itô [5], among others. The initial intuition of Perrin, regarding rotation Brownian motion, was later made rigorous in McKean's 1960 paper [6].

Mathematically [7], a diffusion process X in a manifold M is determined by its initial distribution μ and its infinitesimal generator A, which is a second order differential operator. For instance, if the manifold M is Riemannian, a diffusion process X whose generator is $A = (1/2)\Delta$, half of the Laplacian operator of M, is known as Riemannian Brownian motion.

On the other hand, all real world models depend on parameters which characterise, (for example), time scales, microscopic properties, or effects of the environment. Accordingly, in concrete applications, one is faced with *parameterised diffusions.*

* Postal address: Department of Electrical and Electronic Engineering, Victoria 3010 Australia

** Postal address: Department of Electrical and Electronic Engineering, Victoria 3010 Australia

A parameterised diffusion X in the manifold M is given by a parameter space Θ, here $\Theta \subset \mathbb{R}^p$ for some $p \geq 1$, and a parametric family of initial distributions μ_θ and generators A_θ. That is, by a rule which associates μ_θ and A_θ to each value of the parameter $\theta \in \Theta$. Alternatively, a parameterised diffusion is given by a parametric family of stochastic differential equations, driven by Brownian motion, on the manifold M. The process X then corresponds to weak solutions of these stochastic differential equations — See discussion in Paragraph 2.1 and Appendix A. Both of these approaches (using the generator, or a stochastic differential equation) are concerned with providing a local description of the diffusion X.

This paper is motivated by the following problem. A path of the parameterised diffusion X, in the manifold M, is observed in continuous time. Precisely, the available observation is X_t where t ranges over a finite interval $0 \leq t \leq T$. Based on this observation, estimates of the true value of the parameter θ are to be constructed. The aim is to find estimates which have good asymptotic properties. That is, in the very least, estimates which converge to the true value of θ as $T \rightarrow \infty$. To address this problem, the paper studies the method of maximum likelihood estimation. Its main results are to derive the likelihood function and likelihood equation and prove consistency and asymptotic optimality of maximum likelihood estimates.

In the engineering literature, parameter estimation for diffusions in manifolds has mostly been considered through specific applied problems. These include aeronautics [8, 9] and, more recently, optical communication [10] and robotics [11].

On the other hand, to the authors' knowledge, whether in the engineering or in the mathematics literature, little attention has been devoted to general parameter estimation problems for diffusions in manifolds, (that is, problems involving a general parameterised diffusion on a general differentiable manifold). One exception is the paper by Ng, Caines and Chen [12], concerned with the maximum likelihood method. It derives the likelihood function, but does not study the asymptotic properties of maximum likelihood estimates.

It seems that, at present, there exists no systematic study of the asymptotic properties of maximum likelihood estimation for diffusions in manifolds. The current paper proposes to address precisely this issue.

The paper applies the following methodology. For parameterised diffusions in Euclidean space, the theory of maximum likelihood estimation, based on continuous time observation, is well established. The paper combines existing results for diffusions in Euclidean space with the tools of stochastic differential geometry, in order to generalise them to diffusions in manifolds. The same general approach has been applied to the problem of filtering with observation in a manifold [13–16].

For a complete account of maximum likelihood estimation for scalar diffusions, in continuous time, see the monograph by Kutoyants [17]. For the following, all required background from stochastic differential geometry can be found in [18] or [19].

The paper opens with Section 2, which is concerned with the definition, geometry and ergodicity of parameterised diffusions. Paragraph 2.1 defines a parameterised diffusion X on a manifold M. It shows that X induces a parametric family of probability measures $\{P_\theta; \theta \in \Theta\}$ on the space Ω of continuous paths in M. Here, $\Theta \subset \mathbb{R}^p$ is the parameter space, and $\{P_\theta; \theta \in \Theta\}$ will be called the parametric model. From the outset, it is assumed that X is an elliptic diffusion. This means that X defines a Riemannian metric on the manifold M.

In order to study asymptotic properties of maximum likelihood estimates, (or any other kind of estimates), it is necessary to ensure that X does not explode, (that is, X_t can be observed for all finite time t), and that it is ergodic. Geometric conditions, which guarantee that X does not explode, are given in Appendix A, Proposition 6, using the concept of stochastic completeness. Ergodicity of X is characterised in Paragraph 2.3. These results rely directly on the assumption that X is elliptic.

Paragraph 2.2 develops the stochastic differential geometry of the diffusion X. It introduces the antidevelopment process x of X. While, in general, x is not a diffusion, it has its values in a Euclidean space and can be used to give a simple characterisation of the parametric model $\{P_\theta; \theta \in \Theta\}$. This is stated precisely in Theorem 1, which is of fundamental importance to the whole paper. Note that antidevelopment also played the central role in [16].

Sections 3, 4 and 5 are concerned with maximum likelihood estimation and its asymptotic properties. Section 3 uses Theorem 1 and Girsanov's theorem to derive the likelihood function and likelihood equation for the parametric model $\{P_\theta; \theta \in \Theta\}$, defined in Section 2. The main result of this section is Proposition 1, which gives the likelihood function. The likelihood equation is given in Paragraph 3.2. This equation depends on the length T of the interval of observation, (as explained above, X is only observed over a finite interval). Its solution θ^*_T is the maximum likelihood estimate.

Section 4 proves Propositions 2 and 3. Proposition 2 states that the maximum likelihood estimate θ^*_T is consistent. That is, it converges to the true value of the parameter θ as $T \to \infty$. Proposition 3 states that θ^*_T is asymptotically normal. That is, the difference between θ^*_T and the true value of the parameter θ, is asymptotically distributed according to a normal distribution. This normal distribution has zero mean and its covariance matrix is the inverse of the so called Fisher information matrix. In particular, a byproduct of Proposition 3 is to give the expression of the Fisher information matrix for parameterised diffusions in manifolds.

Section 5 proves the asymptotic optimality of maximum likelihood estimation. A class of estimation methods is introduced, (based on the concept of estimating function as defined by Heyde [20]), of which maximum likelihood is a special case. Propositions 4 and 5 show that, even when other estimation methods lead to estimates which are asymptotically normal, the smallest possible asymptotic covariance matrix is obtained by using maximum likelihood estimation, (recall if A and B are symmetric positive definite matrices, A is said to be smaller than B if $B - A$ is positive definite). Thus, maximum likelihood estimation is asymptotically optimal within the considered class of estimation methods.

Here is a heuristic description of how maximum likelihood works in the current setting. Let X be a parameterised diffusion in the manifold M. In a purely formal way, assume X has a differential dX, such that dX_t can be treated as a tangent vector to M at X_t. The first step is to identify the drift part of dX_t. This should be equal to $D_\theta(X_t)dt$, where D_θ is a vector field on M depending on the parameter θ. Drift represents the "deterministic part" of dX_t. Once it is removed, $dX_t - D_\theta(X_t)dt$ is the "pure diffusion" part. In particular, $dX_t - D_\theta(X_t)dt$ should have zero expectation. The basic idea behind the likelihood equation is to consider for $\theta \in \Theta$ the process $w_T(\theta)$ defined as follows, (here $\langle \cdot, \cdot \rangle$ is the Riemannian metric of M),

$$w_T(\theta) = \int_0^T \langle K, dX_t - D_\theta(X_t)dt \rangle$$

where K is any suitable process such that K_t is a tangent vector to M at X_t. If θ is the true value of the parameter, since $dX_t - D_\theta(X_t)$ has zero expectation, the process $w_T(\theta)$ should be a zero expectation square integrable martingale. Assuming the diffusion X is ergodic, one hopes to obtain an asymptotically normal estimate of the true vaue of θ by solving the equation $(1/\sqrt{T})w_T(\theta) = 0$ for $\theta \in \Theta$. The likelihood equation arises by searching for the process K which gives optimal asymptotic performance, in the sense of having the smallest possible asymptotic covariance matrix — Compare to Section 5.

In the above explanation, notation like dX_t or $D_\theta(X_t)dt$ was not given a precise meaning. The correct definition of dX_t, or of the integral in the above expression for $w_T(\theta)$, is given in Paragraph 2.2. The various ways of defining D_θ are discussed in Section 6. On the whole, the paper carries out in a rigorous mathematical way the heuristic approach just described.

Section 7 discusses the application of maximum likelihood estimation, (precisely, the likelihood equation of Paragraph 3.2), to parameter estimation for diffusions in Lie groups and symmetric spaces. It serves as an example, or case study, allowing the results of Sections 3 and 4 to be discussed in a concrete setting.

2. Parameterised elliptic diffusions

2.1. The parametric model

Let M be a smooth manifold of dimension d and Θ an open subset or $\mathbb{R}^p$. Respectively, M and Θ are the state space and the parameter space.

Observation comes in the form of continuous paths ω, where $\omega(t) \in M$ for $t \geq 0$; (t represents time). The space of such paths is $\Omega = C(\mathbb{R}_+, M)$, which is the sample space. The state at time $t \geq 0$ is the mapping $X_t : \Omega \to M$ where $X_t(\omega) = \omega_t$.

The sample space Ω is considered with the Borel σ-field $\mathcal{F}$ generated by the topology of local uniform convergence. A parametric model associates to each $\theta \in \Theta$ a probability measure P_θ on $\mathcal{F}$. This is such that the observation process $X = \{X_t; t \geq 0\}$ is an elliptic diffusion with values in M. The model $\{P_\theta; \theta \in \Theta\}$ is constructed as follows.

Assume given vector fields $(V_r; r = 1, \dots, v)$ on M along with a smooth function $H : \Theta \times M \to TM$, such that H_θ defined by $H_\theta(x) = H(\theta, x)$ is a also a vector field on M. Consider, for $\theta \in \Theta$ and $x \in M$, the stochastic differential equation,

$$dY_t = H_\theta(Y_t)dt + \sum_{r=1}^{v} V_r(Y_t) \circ dB_t^r \qquad Y_0 = x \tag{1}$$

Here, the unknown process $Y = \{Y_t; t \geq 0\}$ is required to be pathwise continuous with values in M. Moreover, $\circ dB_t^r$ denotes the Stratonovich differential of a standard (unit strength) Brownian motion $(B^r; r = 1, \dots, v)$. The following hypothesis is made,

(H1) For each $\theta \in \Theta$ and $x \in M$, equation (1) has a unique weak solution Y_θ^x.

This means that, (see [7]), it is possible to construct a probability space on which a Brownian motion $(B^r; r = 1, \dots, v)$ and a process Y_θ^x are defined which together satisfy (1). The probability measure induced by Y_θ^x on $\mathcal{F}$, (the distribution of the paths of this process), is denoted P_θ^x.

To specify P_θ, it remains to specify the distribution of X_0. Let $\{\mu_\theta; \theta \in \Theta\}$ be a family of probability measures, on the Borel σ-field $\mathcal{B}(M)$, and define

$$P_\theta(A) = \int P^x_\theta(A)\mu_\theta(dx) \qquad A \in \mathcal{F} \tag{2}$$

Then, the distribution of X_0 is $P_\theta \circ X_0^{-1} = \mu_\theta$. Note that the state X_t can only be observed over a finite time interval $t \leq T$. For $t \geq 0$, let $\mathcal{F}_t = \sigma\{X_s; s \leq t\}$, so $\mathcal{F}_t \subset \mathcal{F}$. In practice, one is really interested in the restriction of each P_θ to $\mathcal{F}_T$.

Throughout the following, the assumption is made that X is an elliptic diffusion. That is, it is assumed the vectors $(V_r(x); r = 1, \ldots, v)$ span the tangent space T_xM at each $x \in M$. This is a natural assumption which guarantees that, in (1), H_θ cannot be separated from the "noise term" by a some linear transformation.

Hypothesis **(H1)** is somewhat strong, as it requires weak solutions of (1) are defined for all $t \geq 0$. In other words, it requires these solutions do not explode. Under the assumption that X is elliptic, sufficient conditions of a geometric nature can be stated which guarantee hypothesis **(H1)** holds. See Proposition 6, Appendix A.

2.2. The geometry of elliptic diffusion

The main result of the current paragraph is Theorem 1, which uses the geometry of equation (1) to give a simple characterisation of the parametric model $\{P_\theta; \theta \in \Theta\}$. This theorem is the very basis for the study of maximum likelihood estimation, carried out in subsequent sections.

Note before going on that the only filtration considered on Ω will be $\{\mathcal{F}_t; t \geq 0\}$. Thus, words like "local martingale" or "Brownian motion" should be taken to imply "with respect to $\{\mathcal{F}_t; t \geq 0\}$". For $\theta \in \Theta$, let A_θ be the differential operator

$$A_\theta f = H_\theta f + \frac{1}{2}\sum_{r=1}^{v} V_r^2 f \tag{3}$$

defined for all smooth function f on M. The probability measure P_θ is uniquely determined by the property that $P_\theta \circ X_0^{-1} = \mu_\theta$ and, for all smooth function f on M,

$$df(X_t) = A_\theta f(X_t)dt + dm^f_t \tag{4}$$

where m^f, is a P_θ-local martingale with $m^f_0 = 0$ — See [7].

The assumption of ellipticity stated in the previous paragraph has the following consequence. There exists, on M, a Riemannian metric $\langle\cdot,\cdot\rangle$ which verifies

$$\langle E, K\rangle = \sum_{r=1}^{v}\langle E, V_r(x)\rangle\langle K, V_r(x)\rangle \qquad E, K \in T_xM \tag{5}$$

With respect to this metric, the gradient and Laplacian of a smooth function f read

$$\text{grad}f = \sum_{r=1}^{v}(V_r f)\, V_r \qquad \Delta f = \sum_{r=1}^{v}\left(V_r^2 - \nabla_{V_r}V_r\right) f \tag{6}$$

Here, ∇ denotes the Levi-Civita connection associated to the metric $\langle\cdot,\cdot\rangle$. With these definitions in mind, it is possible to reformulate condition (4). Define the Itô differential $\langle\text{grad}f, dX_t\rangle$ by

$$df(X_t) = \langle\text{grad}f, dX_t\rangle + \frac{1}{2}\Delta f(X_t)dt \tag{7}$$

Formally, this is the same as a classical Itô formula. Subtracting (4) from (7),

$$\langle \text{grad} f, dX_t \rangle = \left\langle \text{grad} f, H_\theta f(X_t) + \frac{1}{2}\sum_{r=1}^{v} \nabla_{V_r} V_r f(X_t) \right\rangle dt + dm_t^f \tag{8}$$

which is indeed new way of defining the process m^f.

In order to state Theorem 1, it will be necessary to extend the definition of Itô differential $\langle \text{grad} f, dX_t \rangle$ to include $\langle E, dX_t \rangle$ where E is not necessarily the gradient of some smooth function. A vector field above X is a process E with values in TM which is continuous, adapted and such that $E_t \in T_{X_t}M$ for $t \geq 0$. By Whitney's embedding theorem, it is always possible to write, as in [18],

$$E_t = \sum_{\alpha=1}^{n} e_t^\alpha \text{grad} f_\alpha(X_t) \tag{9}$$

where $(e^\alpha; \alpha = 1, \dots, n)$ are real-valued, continuous, adapted and $(f_\alpha; \alpha = 1, \dots, n)$ is an embedding of M in $\mathbb{R}^n$. Now, consistently with (7), let

$$\langle E, dX_t \rangle = \sum_{\alpha=1}^{n} e_t^\alpha \langle \text{grad} f_\alpha, dX_t \rangle \tag{10}$$

This is independent of the chosen functions f_α, since these are required to describe an embedding of M.

Note from (7) and (10),

$$\langle E, dX_t \rangle = \left\langle E_t, H_\theta + \frac{1}{2}\sum_{r=1}^{v} \nabla_{V_r} V_r \right\rangle dt + dm_t^E \qquad dm_t^E = \sum_{\alpha=1}^{n} e_t^\alpha dm_t^{f_\alpha} \tag{11}$$

so that m^E is a P_θ-local martingale with $m_0^E = 0$, for any $\theta \in \Theta$. Recall that the quadratic covariation of m^f, m^k for any smooth functions f, k is given by [7]

$$d\left[m^f, m^k\right]_t = \langle \text{grad} f(X_t), \text{grad} k(X_t) \rangle dt$$

It follows from (11) that, for vector fields E, K above X, the quadratic covariation of m^E, m^K is given by

$$d\left[m^E, m^K\right]_t = \langle E_t, K_t \rangle dt \tag{12}$$

At this point, one is tempted to exploit Lévy's characterisation of Brownian motion by introducing an orthonormal system of vector fields above X. This intuition is key to Theorem 1.

An orthonormal frame above X is a family $(E^i; i = 1, \dots, d)$ of vector fields above X such that, for all $t \geq 0$, $\langle E_t^i, E_t^j \rangle = \delta_{ij}$ for all $t \geq 0$. Here, as usual, δ_{ij} is the Kronecker delta symbol. To construct an orthonormal frame above X, it is possible to use the notion of stochastic parallel transport [18]. Here, it is accepted that one exists, noted $(E^i; i = 1, \dots, d)$ and fixed throughout. Consider the process x with values in $\mathbb{R}^d$ whose components are given by

$$dx_t^i = \langle E^i, dX_t \rangle \qquad i = 1, \dots, d \tag{13}$$

Allowing a minor abuse of terminology, this process x will be called the stochastic antidevelopment of X.

The theorem characterises P_θ using the distribution of x.

Theorem 1. *For $\theta \in \Theta$, the probability measure P_θ on $\mathcal{F}$ is uniquely determined by the property that $P_\theta \circ X_0^{-1} = \mu_\theta$ and*

$$dx^i_t = d^i_t(\theta)dt + d\beta^i_t(\theta) \qquad d^i_t(\theta) = \langle E^i_t, D_\theta \rangle \tag{14}$$

where $(\beta^i(\theta); i = 1, \dots, d)$ is a P_θ-Brownian motion and $D_\theta = H_\theta + (1/2)\sum_{r=1}^{v} \nabla_{V_r} V_r$.

Proof. By (2), $P_\theta \circ X_0^{-1} = \mu_\theta$. By (11) and (13),

$$dx^i_t = d^i_t(\theta)dt + dm^i_t$$

where m^i is a P_θ-local martingale. Moreover, by (12),

$$d\left[m^i, m^j\right]_t = \langle E^i, E^j \rangle dt = \delta_{ij} dt$$

Let $\beta^i(\theta) = m^i$. By Lévy's characterisation of Brownian motion, $(\beta^i(\theta); i = 1, \dots, d)$ is a P_θ-Brownian motion.

Conversely, fix an arbitrary $\theta \in \Theta$. Let P be a probability measure on $\mathcal{F}$ with the property that $P \circ X_0^{-1} = \mu_\theta$ and $(\beta^i(\theta); i = 1, \dots, d)$ is a P-Brownian motion. Note that, for any smooth function f on M,

$$\langle \text{grad} f, dX_t \rangle = \sum_{i=1}^{d} (\text{grad} f)_i(t) dx^i_t \qquad (\text{grad} f)_i(t) = \langle \text{grad} f, E^i_t \rangle$$

This is because $(E^i; i = 1, \dots, d)$ is orthonormal and by the chain rule for Itô differentials. It follows from (14) that

$$\langle \text{grad} f, dX_t \rangle = \left\langle \text{grad} f(X_t), H_\theta + \frac{1}{2}\sum_{r=1}^{v} \nabla_{V_r} V_r \right\rangle + dm_t$$

where $dm_t = \sum_{i=1}^{d} (\text{grad} f)_i(t) d\beta^i_t(\theta)$. Clearly, m is then a P-local martingale with $m_0 = 0$. This shows that P verifies condition (8) and thus the equivalent condition (4). By uniqueness of P_θ under this condition, $P = P_\theta$.

2.3. Ergodic property of the model

The aim of this paper is to study maximum likelihood estimation for the parametric model $\{P_\theta; \theta \in \Theta\}$. Precisely, to describe its asymptotic properties which arise when T is made arbitrarily large.

Typically, the study of asymptotic properties requires that P_θ, (for each $\theta \in \Theta$), should have some form of ergodicity. Precisely, there should exist some probability measure μ^*_θ on $\mathcal{B}(M)$, such that for all smooth function f on M

$$P_\theta - \lim_{T \to \infty} \frac{1}{T} \int_0^T f(X_t)dt = E^*_\theta(f) \tag{15}$$

Here, $P_\theta - \lim$ notes the limit in probability with respect to P_θ and E^*_θ denotes expectation with respect to μ^*_θ on $\mathcal{B}(M)$.

The assumption of ellipticity made in Paragraph 2.1, implies μ^*_θ exists if and only if $\mu^*_\theta(dx) = p_\theta(x)v(dx)$ where v is the Riemannian volume measure on M, (recall M

is equipped with the metric (5)), and p_θ is a smooth strictly positive function on M with [7]

$$\int p_\theta(x)v(dx) = 1 \qquad A^*_\theta p_\theta = 0 \tag{16}$$

Here, A^*_θ is the formal adjoint of A_θ. For any smooth function p on M,

$$A^*_\theta p = -\text{div}(pD_\theta) + \frac{1}{2}\Delta p \tag{17}$$

Recall the Laplacian Δ was defined in (6), Paragraph 2.2, and D_θ was defined in Theorem 1. The divergence of a vector field D on M is

$$\text{div}(D) = \sum_{r=1}^{v}\langle V_r, \nabla_{V_r} D\rangle \tag{18}$$

The following hypothesis is made,

(H2) For $\theta \in \Theta$, there exists a smooth strictly positive function p_θ verifying (16). Moreover, $\mu_\theta(dx) = p_\theta(x)v(dx)$, so that the observation process X is P_θ-stationary.

Note that uniqueness of p_θ for this hypothesis follows from (15). Hypothesis **(H2)** is verified whenever M is compact. In general, even if M is not compact, assume a smooth function $U : \Theta \times M \to \mathbb{R}$ is given, such that

$$Z_\theta = \int \exp(-2U_\theta(x))v(dx) < \infty \tag{19}$$

where $U(\theta, x) = U_\theta(x)$. Then [21], if $D_\theta = -\text{grad}(U_\theta)$, hypothesis **(H2)** holds for

$$p_\theta = Z^{-1}_\theta \exp(-2U_\theta) \tag{20}$$

3. Maximum likelihood estimation

Maximum likelihood estimation for the parametric model $\{P_\theta; \theta \in \Theta\}$ proceeds along the following lines — Compare to [17].

Fix some $\rho \in \Theta$. Assume it can be shown that, for $\theta \in \Theta$ and $T \geq 0$,

$$\left.\frac{dP_\theta}{dP_\rho}\right|_T = L_T(\theta) \qquad L_T(\theta) > 0 \tag{21}$$

where the subscript T on the left hand side denotes restriction of P_θ and P_ρ to $\mathcal{F}_T$. Then, the maximum likelihood estimate θ^*_T of θ is defined to be any $\mathcal{F}_T$-measurable random variable with values in Θ such that

$$L_T(\theta^*_T) = \sup_{\theta\in\Theta} L_T(\theta) \tag{22}$$

Note that $L_T(\theta)$ is a random function of $\theta \in \Theta$, known as the likelihood function. This is worth emphasising since, in definition (21), dependence on the observation $\omega \in \Omega$ was suppressed.

An alternative definition of θ^*_T requires an additional differentiability property of $L_T(\theta)$. Consider the log-likelihood function $\ell_T(\theta) = \log[L_T(\theta)]$. The maximum likelihood estimate θ^*_T of θ may be defined to be a $\mathcal{F}_T$-measurable random variable with values in Θ solving the equation

$$\partial \ell_T(\theta^*_T) = 0 \tag{23}$$

where ∂ denotes the derivative with respect to θ.

Recall $\Theta \subset \mathbb{R}^p$, so that (23) is a system of p equations in p unknowns. Definitions (22) and (23) are not equivalent. However, if θ^*_T verifies (22) and $L_T(\theta)$ is differentiable, then θ^*_T also verifies (23).

3.1. The likelihood ratio

This paragraph is concerned with the existence of a likelihood function $L_T(\theta)$, as in (21).

The main result is Proposition 1 below, which refers to the following hypothesis,

(H3) For $\theta', \theta \in \Theta$, it holds that $\sup_{x\in M} \|H_{\theta'}(x) - H_\theta(x)\| < +\infty$.

Here, $\|\cdot\|$ denotes Riemannian length. The idea of Proposition 1 will be to apply Girsanov's theorem to Theorem 1.

Proposition 1. *Assume hypotheses **(H1-H3)** hold. For $\theta', \theta \in \Theta$ and $T \geq 0$,*

$$\left.\frac{dP_{\theta'}}{dP_\theta}\right|_T = L_T(\theta', \theta) \tag{24}$$

where the likelihood ratio $L_T(\theta', \theta)$ is given by

$$L_T(\theta', \theta) = \frac{p_{\theta'}}{p_\theta}(X_0) \exp\left(\int_0^T \left(\delta_t(\theta', \theta), d\beta_t(\theta)\right) dt - \frac{1}{2}\int_0^T \left|\delta_t(\theta', \theta)\right|^2 dt \right) \tag{25}$$

Here, $(\cdot, \cdot)$ and $|\cdot|$ denote Euclidean scalar product and norm on $\mathbb{R}^d$.

Equivalently,

$$L_T(\theta', \theta) = \frac{p_{\theta'}}{p_\theta}(X_0) \exp\left(\int_0^T \langle \Delta(\theta', \theta), dX_t - D_\theta(X_t)dt\rangle - \frac{1}{2}\int_0^T \left\|\Delta(\theta', \theta)(X_t)\right\|^2 dt \right) \tag{26}$$

Here, $\Delta(\theta', \theta) = H_{\theta'} - H_\theta$ and $\delta(\theta', \theta)$ has its values in $\mathbb{R}^d$ with $\delta^i_t(\theta', \theta) = \langle E^i_t, \Delta(\theta', \theta)\rangle$ for $i = 1, \ldots, d$.

Proof. Hypothesis **(H1)** guarantees the model $\{P_\theta; \theta \in \Theta\}$ is well defined. Hypothesis **(H2)** allows division by p_θ.

First, it is proved that (25) and (26) are equivalent. Recall the definition of $d\beta_t(\theta)$ in (14), $d\beta_t(\theta) = dx_t - d_t(\theta)dt$.

Starting from (26), note that

$$\langle \Delta(\theta', \theta), dX_t - D_\theta(X_t)dt\rangle = (\delta_t(\theta', \theta), dx_t - d_t(\theta)dt)$$

This is because $(E^i; i = 1, \ldots, d)$ is orthonormal and by the chain rule for Itô differentials, (compare to the proof of Theorem 1). It follows that

$$\langle \Delta(\theta', \theta), dX_t - D_\theta(X_t)dt\rangle = (\delta_t(\theta', \theta), d\beta_t(\theta))$$

Similarly, due to the fact that $(E^i; i = 1, \dots, d)$ is orthonormal,

$$\|\Delta(\theta', \theta)(X_t)\|^2 \, dt = |\delta_t(\theta', \theta)|^2 \, dt$$

From the last two equalities, the expressions under the exponential in (25) and (26) are the same.

It remains to prove (24). Note from (14),

$$d\beta_t(\theta') = d\beta_t(\theta) - \delta(\theta', \theta) dt \tag{27}$$

By Theorem 1, $\beta(\theta)$ is a P_θ-Brownian motion.

Under hypothesis **(H3)**, the process $L(\theta', \theta)$ given by (25) is a P_θ-martingale. Thus, a probability measure P on $\mathcal{F}$ can be defined by the change of measure formula

$$\left. \frac{dP}{dP_\theta} \right|_T = L_T(\theta', \theta)$$

By Girsanov's theorem, $\beta(\theta')$ is then a P-Brownian motion. Since $L_0(\theta', \theta)$ is equal to $(p_{\theta'}/p_\theta)(X_0)$, it follows $P \circ X_0^{-1} = (p_{\theta'}/p_\theta)\mu_\theta = \mu_{\theta'}$. Then, $P = P_{\theta'}$, by the uniqueness statement in Theorem 1.

The existence of a likelihood function, as in (21) follows immediately from this proposition. One simply needs to choose a reference probability P_ρ, where $\rho \in \Theta$, and set $L_T(\theta) = L_T(\rho, \theta)$. Note finally that hypothesis **(H3)** was only used in showing $L(\theta', \theta)$ is a P_θ-martingale. For this purpose, it can be replaced by weaker hypotheses such as Novikov's condition [22].

3.2. The likelihood equation

This paragraph is concerned with equation (23), which will be called the likelihood equation. The main objective is to write this equation down using Proposition 1. Note that (25), of Proposition 1, immediately yields

$$\ell_T(\theta) = \log\left(\frac{p_\theta}{p_\rho}(X_0)\right) + \int_0^T (\delta_t(\theta, \rho), d\beta_t(\rho)) - \frac{1}{2}\int_0^T |\delta_t(\theta, \rho)|^2 \, dt \tag{28}$$

Assume it is possible to differentiate under the integrals, stochastic or ordinary. Replacing the definitions of $d_t(\theta)$ and $\delta_t(\theta, \rho)$, (see Theorem 1 and Proposition 1), it follows by a straightforward calculation

$$\partial \ell_T(\theta) = \partial \log p_\theta(X_0) + \int_0^T (\partial d_t(\theta), d\beta_t(\rho)) \tag{29}$$

Or, directly in terms of X,

$$\partial \ell_T(\theta) = \partial \log p_\theta(X_0) + \int_0^T \langle \partial D_\theta, dX_t - D_\theta(X_t) dt \rangle \tag{30}$$

In (29) and (30), the derivatives $\partial d(\theta)$ and ∂D_θ are integrated component by component. Recall that θ denotes an element in $\mathbb{R}^p$, say $\theta = (\theta^a; a = 1, \dots, p)$. The components of $\partial d(\theta)$ are the partial derivatives $\partial d(\theta)/\partial \theta^a$. These are processes with values in $\mathbb{R}^d$, so they can be integrated against $d\beta(\rho)$. Similarly, the components of ∂D_θ are the partial derivatives $\partial D_\theta / \partial \theta^a$. These are vector fields above X and can be integrated against dX, according to (11). Now, $\partial \ell_T(\theta)$ is a random function of $\theta \in \Theta$ and with range in $\mathbb{R}^p$; (it is known as the score function). Based on (30), the likelihood equation (23) takes the form

$$\partial \log p_{\theta^*_T}(X_0) + \int_0^T \langle \partial D_{\theta^*_T}, dX_t - D_{\theta^*_T}(X_t)dt \rangle = 0 \tag{31}$$

For now, this rests on the assumption that it is possible to differentiate under the integrals, in particular the stochastic integral, in (28).

In [23], Karandikar gives surprisingly weak conditions which guarantee this assumption holds. Theorem 5 on page 124 of [23], applied to the current context, shows that (29) is correct as soon as hypothesis **(H1)** is verified and the function $H : \Theta \times M \rightarrow TM$ possesses locally Lipschitz partial derivatives $(\partial H / \partial \theta^a; a = 1, \dots, p)$. In Paragraph 2.1, H was introduced as a smooth function, so that it does have the property just mentioned.

Recall, finally, that (30) is equivalent to (29).

4. Asymptotic properties of maximum likelihood

Maximum likelihood estimation is often used for its good asymptotic properties.

This section is concerned with the properties of consistency and asymptotic normality, which it respectively states in Propositions 2 and 3 of Paragraphs 4.1 and 4.2 below.

4.1. Consistency of $\boldsymbol{\theta^*_T}$

Roughly, consistency means that the maximum likelihood estimate θ^*_T converges in probability to the "true value" of the parameter θ, as $T \rightarrow \infty$. This somewhat confusing statement translates to the following mathematical condition. For $\theta \in \Theta$, any random variables $\{\theta^*_T; T \geq 0\}$, defined by (22), verify

$$P_\theta - \lim_{T \rightarrow \infty} \theta^*_T = \theta \tag{32}$$

Note that, in general, the supremum in (22) may not be achieved. In this event, set $\theta^*_T = \infty$. Condition (32) is understood with this convention.

Proposition 2 will require the following identifiability hypothesis,

(H4) For any $\theta, \theta' \in \Theta$, the identity $H_\theta(x) = H_{\theta'}(x)$ for all $x \in M$ implies $\theta = \theta'$.

Proposition 2. *Assume hypotheses **(H1-H4)** hold and Θ is bounded. If there exist constants $\beta > 1, \gamma > p$ and $C > 0$ such that*

$$E^*_\theta \|\Delta(\theta'', \theta')\|^\beta \leq C|\theta'' - \theta'|^\gamma \qquad E^*_\theta \left\{ \|\Delta(\theta'', \theta)\|^2 - \|\Delta(\theta', \theta)\|^2 \right\}^\beta \leq C|\theta' - \theta''|^\gamma \tag{33}$$

for all $\theta, \theta', \theta'' \in \Theta$, then (32) holds for all $\theta \in \Theta$.

Proof. Hypotheses **(H1-H3)** guarantee that Proposition 1 holds.

Recall, from (21) and (24), that $L_T(\theta') = L_T(\theta', \theta) L_T(\theta)$. It follows that θ^*_T, given by (22), satisfies

$$L_T(\theta^*_T, \theta) = \sup_{\theta' \in \Theta} L_T(\theta', \theta) \tag{34}$$

Fix an arbitrary $\theta \in \Theta$ and consider the function $g : \Theta \to \mathbb{R}_+$,

$$g(\theta') = -\frac{1}{2} E^*_\theta \|\Delta(\theta', \theta)\|^2 \qquad \theta' \in \Theta \tag{35}$$

It follows from hypotheses **(H4)** that g has a unique global maximum at θ.

For $T \geq 1$, consider the random function $g_T : \Theta \to \mathbb{R}_+$ where $g_T(\theta') = T^{-1}\ell(\theta', \theta)$ for $\theta' \in \Theta$, where $\ell_T(\theta', \theta) = \log[L_T(\theta', \theta)]$. In other words, as in (28),

$$g_T(\theta') = \frac{1}{T} \log \left(\frac{p_{\theta'}}{p_\theta} (X_0) \right) + \frac{1}{T} \int_0^T (\delta_t(\theta', \theta), d\beta_t(\theta)) - \frac{1}{2T} \int_0^T |\delta_t(\theta', \theta)|^2 \, dt \tag{36}$$

Let Q_T and Q denote, respectively, the probability measures on the (Borel σ-field of the) the space $C(\Theta, \mathbb{R})$, which are the images of P_θ with respect to g_T and g. Using the Kolmogorov-Chentsov tightness condition, it is now shown $Q_T \Rightarrow Q$ as $T \to \infty$; ($\Rightarrow$ denotes weak convergence of probability measures — See [24], page 313).

Note the first term on the right hand side of (36) converges to zero, identically on Ω, as $T \to \infty$. Therefore, it will simply be ignored in the remainder of the proof.

Let $I_T(\theta')$ denote the second term and $V_T(\theta)$ denote the third term, on the right hand side of (36). Theorem 1 states $\beta(\theta)$ is a P_θ-Brownian motion. It follows that

$$E_\theta |I_T(\theta')|^2 = \frac{1}{T^2} \int_0^T E_\theta \|\Delta(\theta', \theta)(X_t)\|^2 dt \leq \frac{1}{T} \sup_{x \in M} \|\Delta(\theta', \theta)(x)\|^2$$

where E_θ denotes expectation with respect to P_θ. Hypothesis **(H3)** states the supremum appearing here is finite. Thus, $P_\theta - \lim_{T \to \infty} I_T(\theta') = 0$. Applying (15) to $V_T(\theta)$, it follows that

$$P_\theta - \lim_{T \to \infty} g_T(\theta') = g(\theta') \tag{37}$$

This shows that finite dimensional projections of Q_T converge weakly to finite dimensional projections of Q. The Kolmogorov-Chentsov condition follows from (33). Note first that, for $\theta', \theta'' \in \Theta$,

$$\begin{aligned} E_\theta |I_T(\theta'') - I_T(\theta')|^\beta & \leq C_\beta E_\theta \left(\tfrac{1}{T} \textstyle\int_0^T \|\Delta(\theta'', \theta)\|^2 (X_t) dt \right)^{\beta/2} \\ & \leq C_\beta E_\theta \left(\tfrac{1}{T} \textstyle\int_0^T \|\Delta(\theta'', \theta)\|^\beta (X_t) dt \right) \end{aligned}$$

The first inequality is the Burkholder-Davis-Gundi inequality, where C_β is a universal constant. The second inequality follows from Jensen's inequality. When combined with hypothesis **(H2)**, which states X is P_θ stationary, and (33), this yields

$$E_\theta |I_T(\theta'') - I_T(\theta')|^\beta \leq C |\theta'' - \theta'|^\gamma \tag{38}$$

Similarly, it is possible to show

$$E_\theta |V_T(\theta'') - I_T(\theta')|^\beta \leq C |\theta'' - \theta'|^\gamma \tag{39}$$

This is found using the same steps as for (33), but using Hölder's inequality for ordinary integrals instead of the Burkholder-Davis-Gundi inequality for stochastic integrals.

Combining (38) and (39), it follows the Kolmogorov-Chentsov tightness condition is verified by the probability measures $\{Q_T; T \geq 0\}$. Therefore, $Q_T \Rightarrow Q$. With this result, (32) is proved by a classical reasoning. Let $U \subset \Theta$ be any neighborhood of θ. From (34),

$$P_\theta(\theta^*_T \notin U) = P_\theta\left(\sup_{\theta' \notin U} g_T(\theta') > \sup_{\theta' \in U} g_T(\theta') \right)$$

Since $Q_T \Rightarrow Q$ and Θ is bounded, the right hand side converges to zero as $T \to \infty$. Indeed, under Q the supremum over Θ can only occur at θ. By taking U arbitrarily small, it follows that (32) holds.

4.2. Asymptotic normality of θ^*_T

The property of asymptotic normality states that, for any $\theta \in \Theta$, the distribution of $\theta^*_T - \theta$, with respect to the probaility measure P_θ on $\mathcal{F}$, is asymptotically normal. This is proved in Proposition 3 below, which refers to the following hypothesis

(H5) For each $\theta \in \Theta$, the "Fisher information matrix" $I(\theta)$ is invertible. Here,

$$I_{ab}(\theta) = E^*_\theta \left\langle \partial H_\theta/\partial\theta^a, \partial H_\theta/\partial\theta^b \right\rangle \qquad a, b = 1, \dots, p \tag{40}$$

In the following, whenever Z is a $\mathcal{F}$-measurable random variable, $\mathcal{L}_\theta\{Z\}$ denotes its distribution with respect to P_θ. That is, $\mathcal{L}_\theta\{Z\} = P_\theta \circ Z^{-1}$.

Proposition 3. *Assume hypothesis **(H5)** holds and the conditions of Proposition 2 are verified. If $E^*_\theta\left(\partial^2\right) < +\infty$, where $\partial^2 : M \to \mathbb{R}$ is given by*

$$\partial^2(x) = \sup_{\theta\in\Theta} \sum_{a,b=1}^{p} \left\| \frac{\partial^2 D_\theta(x)}{\partial\theta^a \partial\theta^b} \right\|^2 \tag{41}$$

then,

$$\mathcal{L}_\theta\{\sqrt{T}(\theta^*_T - \theta)\} \Rightarrow N\left(I^{-1}(\theta)\right) \tag{42}$$

Here, $N(C)$ denotes a normal distribution with zero mean and covariance matrix C.

Proof. The notation from the proof of Proposition 2 is here maintained. Fix an arbitrary $\theta \in \Theta$ and a convex neighborhood $U \subset \Theta$ of θ. By Proposition 2, $\theta^*_T \in U$ with high probability as $T \to \infty$. Accordingly, in the following, all random variables are restricted to this event. For $\theta' \in U$, let

$$z^a_T(\theta') = \sqrt{T}\partial_a g_T(\theta') \qquad a = 1, \dots, p \tag{43}$$

Here, ∂_a denotes the partial derivative of g_T with respect to its a-th argument. The notation $z_T(\theta')$ will stand for the vector $(z^a_T(\theta'); a = 1, \dots, p)$. Differentiating under the integrals in (36) is justified by the results of [23] — See the argument after (31) in Paragraph 3.2. This yields, by direct calculation,

$$z^a_T(\theta') = \frac{1}{\sqrt{T}} \int_0^T (\partial_a d_t(\theta'), d\beta_t(\theta) - \delta_t(\theta',\theta)dt) \tag{44}$$

Note that the first term of (36) was ignored, just like in the proof of Proposition 2. In particular, putting $\theta' = \theta$,

$$z^a_T(\theta) = \frac{1}{\sqrt{T}} \int_0^T \left(\partial_a d_t(\theta), d\beta_t(\theta)\right)$$

Theorem 1 states $\beta(\theta)$ is a P_θ-Brownian motion. By the central limit theorem for stochastic integrals [17]

$$\mathcal{L}_\theta\{z_T(\theta)\} \Rightarrow N(I(\theta)) \tag{45}$$

The asymptotic covariance matrix $I(\theta)$ follows from (15),

$$P_\theta - \lim_{T\to\infty} \frac{1}{T} \int_0^T \left(\partial_a d_t(\theta), \partial_b d_t(\theta)\right) dt = E^*_\theta \langle \partial_a D_\theta, \partial_b D_\theta \rangle = I_{ab}(\theta) \tag{46}$$

Since $\theta^*_T \in U$, it is possible to apply the mean value theorem

$$z^a_T(\theta^*_T) = z^a_T(\theta) + \sum_{b=1}^p \partial_b z^a_T(\theta''_a)(\theta^*_T - \theta)^b \tag{47}$$

The superscript b denotes the b-th component of $\theta^*_T - \theta$. Here, $\theta''_a \in U$ lies on the segment connecting θ to θ^*_T. Note that, by (34), the left hand side is zero. To prove (42) holds, it will be enough to prove

$$P_\theta - \lim_{T\to\infty} \frac{1}{\sqrt{T}} \partial_b z^a_T(\theta''_a) = -I_{ab}(\theta) \tag{48}$$

Indeed, hypothesis **(H5)** then guarantees it is possible to multiply either side of (47) by $I^{-1}_{ca}(\theta)$ and sum over a. Let θ' be any one of the θ''_a, where $a = 1, \dots, p$, so that $|\theta' - \theta| < |\theta^*_T - \theta|$ — Here, $|\cdot|$ denotes Euclidean norm on $\mathbb{R}^p$. Derivation under the integrals in (44) gives, ($\partial_{ab} = \partial_a \partial_b$ denote mixed second derivatives),

$$\begin{aligned} \tfrac{1}{\sqrt{T}} \partial_b z^a_T(\theta') \quad &= \tfrac{1}{T} \textstyle\int_0^T \left(\partial_{ab} d_t(\theta'), d\beta_t(\theta)\right) \\ &- \tfrac{1}{T} \textstyle\int_0^T \left(\partial_{ab} d_t(\theta'), \delta_t(\theta', \theta)\right) - \tfrac{1}{T} \textstyle\int_0^T \left(\partial_a d_t(\theta'), \partial_b d_t(\theta')\right) dt \end{aligned}$$

The variance of the first term on the right hand side is,

$$\frac{1}{T^2} \int_0^T E_\theta \|\partial_{ab} D_{\theta'}(X_t)\|^2 dt \leq \frac{1}{T^2} \int_0^T E_\theta \left(\partial^2(X_t)\right) dt \leq \frac{1}{T} E^*_\theta \left(\partial^2\right)$$

Since $E^*_\theta \left(\partial^2\right) < +\infty$, this converges to zero as $T \to \infty$. For the second term, note

$$\frac{1}{T} \int_0^T \left(\partial_{ab} d_t(\theta'), \delta_t(\theta', \theta)\right) dt = \frac{1}{T} \int_0^T \langle \partial_{ab} D_{\theta'}(X_t), \Delta(\theta', \theta)(X_t) \rangle \, dt \tag{49}$$

Using Proposition 2, and $\Delta(\theta, \theta) = 0$, this can be shown to converge to zero in probability as $T \to \infty$. For the third term, note similarly

$$\frac{1}{T} \int_0^T \left(\partial_a d_t(\theta'), \partial_b d_t(\theta')\right) dt = \frac{1}{T} \int_0^T \langle \partial_a D_{\theta'}(X_t), \partial_b D_{\theta'}(X_t) \rangle \, dt \tag{50}$$

The condition $E^*_\theta\left(\partial^2\right) < +\infty$ implies, for some constant C depending on U,

$$E_\theta\left|\langle\partial_a D_{\theta'}(X_t), \partial_b D_{\theta'}(X_t)\rangle - \langle\partial_a D_\theta(X_t), \partial_b D_\theta(X_t)\rangle\right| \leq C E_\theta\left|\theta^*_T - \theta\right|$$

The expectation on the right hand side is finite, since $U \subset \Theta$ is bounded. Proposition 2, using bounded convergence, implies this expectation converges to zero as $T \to \infty$. The required (48) now follows from (46) and (50).

5. Optimality of maximum likelihood estimation

The previous section established consistency and asymptotic normality of maximum likelihood estimation. Here, a further property of *asymptotic optimality* is considered. Precisely, the aim is to show that maximum likelihood estimation provides optimal asymptotic performance within a certain class of other estimation methods.

The main results will be stated in Propositions 4 and 5. These are concerned with a general setting, which is now described.

Consider a slightly modified definition of the maximum likelihood estimate θ^*_T. Recall (44), which can be used to define a random function $z_T : \Theta \to \mathbb{R}^p$. Rewrite (44) using (27) from the proof of Proposition 1. This gives

$$z^a_T(\theta') = \frac{1}{\sqrt{T}}\int_0^T (\partial_a d_t(\theta'), d\beta_t(\theta')) \qquad a = 1, \ldots, p \tag{51}$$

For $\theta' \in \Theta$, the notations $z_T(\theta')$ stands for the vector $(z^a_T(\theta'); a = 1, \ldots, p)$. The results of Propositions 2 and 3 continue to hold, if θ^*_T is defined for $T \geq 0$ as any $\mathcal{F}_T$-measurable random variable with values in Θ and such that $z_T(\theta^*_T) = 0$. This can be seen by going over the proofs of these two propositions, step by step.

Thus, with only a slight abuse of terminology, it is possible to accept that maximum likelihood estimation consists in finding an $\mathcal{F}_T$-measurable root θ^*_T of the random function z_T defined in (51). It is then natural to consider a class of estimation methods defined in a similar way, as follows.

Let $K^a : \Theta \times M \to TM$, where $a = 1, \ldots, p$, be smooth functions such that $K^a_{\theta'}$ defined by $K^a_{\theta'}(x) = K^a(\theta', x)$ are vector fields on M. For $T \geq 0$, define a random function $w_T : \Theta \to \mathbb{R}^p$ by

$$w^a_T(\theta') = \frac{1}{\sqrt{T}}\int_0^T (k^a_t(\theta'), d\beta_t(\theta')) \qquad a = 1, \ldots, p \tag{52}$$

where $k^a_t(\theta')$ is the process with values in $\mathbb{R}^d$ whose components are $\langle E^i_t, K^a_{\theta'}\rangle$, for $i = 1, \ldots, d$. Let $w_T(\theta')$ denote the vector $(w^a_T(\theta'); a = 1, \ldots, p)$ and ρ^*_T be any $\mathcal{F}_T$ measurable random variable with values in Θ and such that $w_T(\rho^*_T) = 0$.

Now, the definition of θ^*_T appears as a special case of the definition of ρ^*_T. Indeed, (51) results from (52) when $K^a_{\theta'} = \partial_a D_{\theta'}$. In light of this observation, Proposition 2 shows that it is possible to choose $K^a_{\theta'}$ so that

$$P_\theta - \lim_{T\to\infty} \rho^*_T = \theta \tag{53}$$

for each $\theta \in \Theta$. Propositions 4 and 5 compare the asymptotic performance of θ^*_T to that of ρ^*_T. Of course, there is no point in this comparison unless ρ^*_T verifies (53).

Proposition 4. *For $\theta \in \Theta$, let $J(\theta), \bar{J}(\theta)$ denote the $p \times p$ matrices with elements*

$$J_{ab}(\theta) = E^*_\theta \langle K^a_\theta, K^b_\theta \rangle \qquad \bar{J}_{ab}(\theta) = E^*_\theta \langle K^a_\theta, \partial_b H_\theta \rangle \tag{54}$$

Assume that $\bar{J}(\theta)$ is invertible and let

$$C(\theta) = \left(\bar{J}^{-1}(\theta)\right) \left(J(\theta)\right) \left(\bar{J}^{-1}(\theta)\right)^\dagger \tag{55}$$

where † denotes the transpose. Assume also that

$$E^*_\theta \left(\sup_{\theta \in \Theta} \sum_{a,b=1}^{p} \|\partial_b K^a\|^2 \right) < +\infty \tag{56}$$

If Θ is bounded and convex and (53) holds, then for $\theta \in \Theta$

$$\mathcal{L}_\theta\{\sqrt{T}(\rho^*_T - \theta)\} \Rightarrow N\left(C(\theta)\right) \tag{57}$$

Proof. The proof closely mirrors that of Proposition 3. Fix an arbitrary $\theta \in \Theta$. By Theorem 1, $\beta(\theta)$ is a P_θ-Brownian motion. By the central limit theorem for stochastic integrals, (again, see [17])

$$\mathcal{L}_\theta\{w_T(\theta)\} \Rightarrow N(J(\theta)) \tag{58}$$

The asymptotic covariance $J(\theta)$ follows from (15),

$$P_\theta - \lim_{T \to \infty} \frac{1}{T} \int_0^T \left(k^a_t(\theta), k^b_t(\theta)\right) dt = E^*_\theta \langle K^a_\theta, K^b_\theta \rangle = J_{ab}(\theta)$$

Since Θ is convex and ρ^*_T is well defined, for sufficiently large $T \geq 0$, it is possible to apply the mean value theorem, (compare to (47) in the proof of Proposition 3),

$$w^a_T(\rho^*_T) = w^a_T(\theta) + \sum_{b=1}^{p} \partial_b w^a_T(\rho''_a)(\rho^*_T - \theta)^b \tag{59}$$

where ρ''_a lies on the segment connecting θ to ρ^*_T. By definition of ρ^*_T, the left hand side is zero. To prove (57), it is enough to prove

$$P_\theta - \lim_{T \to \infty} \frac{1}{\sqrt{T}} \partial_b w^a_T(\rho''_a) = -\bar{J}_{ab}(\theta) \tag{60}$$

To do so, let θ' denote any one of the ρ''_a, where $a = 1, \ldots, p$. Derivation under the integral in (52) gives,

$$\begin{aligned} \tfrac{1}{\sqrt{T}} \partial_b w^a_T(\theta') &= \tfrac{1}{T} \textstyle\int_0^T \left(\partial_b k^a_t(\theta'), d\beta_t(\theta)\right) \\ &- \tfrac{1}{T} \textstyle\int_0^T \left(\partial_b k^a_t(\theta'), \delta_t(\theta', \theta)\right) - \tfrac{1}{T} \textstyle\int_0^T \left(k^a_t(\theta'), \partial_b d_t(\theta')\right) dt \end{aligned}$$

Now, (60) can be proved using (53) and (56). This is done following exactly the same steps as in the proof of Proposition 3.

The following proposition states the asymptotic covariance $I^{-1}(\theta)$ obtained in (42) of Proposition 3 is smaller than any covariance matrix $C(\theta)$ arising in (57) of the previous Proposition 4. In other words, the maximum likelihood estimate θ^*_T has optimal asymptotic performance among all estimates of the form ρ^*_T defined here.

Proposition 5. *Assume the conditions of Propositions 3 and 4 hold. For every $\theta \in \Theta$, the matrix $C(\theta) - I^{-1}(\theta)$ is positive defintie.*

Proof. The proof follows the classical reasoning of Rao, page 327 of [25]. The following matrix is clearly positive definite

$$\begin{pmatrix} J(\theta) & \bar{J}(\theta) \\ \bar{J}^{\dagger}(\theta) & I(\theta) \end{pmatrix}$$

By Rao's reasoning, it follows

$$J(\theta) - \bar{J}(\theta) I^{-1}(\theta) \bar{J}^{\dagger}(\theta)$$

is also positive definite. This is equivalent to the proposition, as $\bar{J}(\theta)$ is invertible.

6. The notion of drift and the Le Jan-Watanabe connection

This section provides some general remarks, which are helpful in interpreting and implementing the estimation methods studied above, mainly maximum likelihood estimation.

Two fundamental questions are discussed. First, how to identify the drift part of the observation process X? Second, how to compute numerically the maximum likelihood estimate θ^*_T? The discussion of the first question, due to the very nature of this question, is rather informal and aimed at building intuition.

The first question underlies the class of estimation methods studied in Section 5. Roughly, by removing from X its drift part, a new object is obtained which is the "pure diffusion" or Brownian part. At least in principle, setting a normalised version of the Brownian part to zero yields an asymptotically normal estimate of the parameter θ.

Theorem 1 of Paragraph 2.2 suggests the drift part of dX_t is $D_\theta(X_t)dt$. Formally, the theorem states that the coordinates of $dX_t - D_\theta(X_t)dt$ in the orthonormal frame $(E^i; i = 1, \dots, d)$ are $d\beta^i_t(\theta)$, where $\beta(\theta)$ is a P_θ-Brownian motion. The theorem makes the even stronger statement that this property uniquely determines P_θ.

This argument leads to a straightforward interpretation of the random functions w_T, defined in (52) of Section 5. Indeed, $w_T(\theta')$ can be written directly in terms of X, (compare to (30) in Paragraph 3.2),

$$w^a_T(\theta') = \frac{1}{\sqrt{T}} \int_0^T \langle K^a_{\theta'}, dX_t - D_{\theta'}(X_t)dt \rangle \qquad a = 1, \dots, p \tag{61}$$

When $\theta' = \theta$, the expression $dX_t - D_{\theta'}(X_t)dt$ appearing here is the Brownian part of dX_t, so w_T is asymptotically normal. The choice of the functions K^a is a just a choice of normalisation determining the asymptotic covariance — See the proof of Proposition 4.

The statement that the drift part of dX_t is $D_\theta(X_t)dt$ seems counterintuitive in view of (1) and (3). Looking at (3), for example, the first order part of the operator A_θ is the vector field H_θ. In the case of scalar or vector diffusions, one is systematically used to identifying drift from the first order part of the infinitesimal generator, (this being A_θ at present).

Note also that the object of interest, as far as the parameter θ is concerned, in Paragraph 2.1, is H_θ. Indeed, X depends on θ only through this function.

Compare now the following identities, satisfied by H_θ and D_θ,

$$H_\theta = A_\theta - \frac{1}{2}\sum_{r=1}^{v} V_r^2 \qquad D_\theta = A_\theta - \frac{1}{2}\Delta \tag{62}$$

where Δ is the Laplacian of M, defined in (6). The "advantage" of D_θ over H_θ is thus purely mathematical. Precisely, using only the Riemannian geometry of M with metric (5), it is possible to give an intrinsic definition of D_θ, but not of H_θ.

To get an intrinsic definition of H_θ, it is necessary to introduce a new geometric construction, the Le Jan-Watanabe connection. This is the affine connection $\bar{\nabla}$ on the tangent bundle of M, defined by the following identity

$$\bar{\nabla}_K E(x) = \sum_{r=1}^{v} K\langle E, V_r\rangle V_r(x) \tag{63}$$

for each $K \in T_xM$ and vector field E on M. Note $K\langle E, V_r\rangle$ is the derivative along the vector K of the function $\langle E, V_r\rangle$. A more detailed account can be found in [16].

Recall now the definition of the Hessian of a smooth function f on M, with respect to the Levi-Civita connection ∇ or the Le Jan-Watanabe connection $\bar{\nabla}$,

$$\nabla^2 f(K,E) = KEf - \nabla_K Ef \qquad \bar{\nabla}^2 f(K,E) = KEf - \bar{\nabla}_K Ef \tag{64}$$

where K, E are vector fields on M. These expressions follow from the standard definition of the Hessian tensor [19]. It is possible to show then that

$$\mathrm{Tr}\nabla^2 f = \Delta f \qquad \mathrm{Tr}\bar{\nabla}^2 f = \sum_{r=1}^{v} V_r^2 f \tag{65}$$

Where Tr indicates the trace. The first identity is a usual definition of the Laplacian. The second follows from

$$\mathrm{Tr}\bar{\nabla}^2 f = \sum_{r=1}^{v} \bar{\nabla}^2 f(V_r, V_r) = \sum_{r=1}^{v} \left(V_r^2 - \bar{\nabla}_{V_r} V_r\right) f$$

However, the last term on the right here is zero, as shown in [16].

Now, (65) provides an intrinsic definition of H_θ. Precisely,

$$H_\theta = A_\theta - \frac{1}{2}\mathrm{Tr}\bar{\nabla}^2 \tag{66}$$

Based on this new definition of H_θ, it is possible to rewrite (61) in a way that completely bypasses $D_{\theta'}$. This is done by introducing a new Itô differential instead of the one given by (7)–(10). Let $(-)\langle \mathrm{grad} f, dX_t\rangle$ be defined by

$$df(X_t) = (-)\langle \mathrm{grad} f, dX_t\rangle + \frac{1}{2}\mathrm{Tr}\bar{\nabla}^2 f(X_t)dt \tag{67}$$

for any smooth function f on M. This extends to vector fields E above X, so that

$$(-)\langle E, dX_t\rangle = \langle E_t, H_\theta\rangle dt + dm_t^E \tag{68}$$

where m^E is the same process as in (11). With this definition of the Itô differential, it is straightforward to verify that (61) is equivalent to

$$w^a_T(\theta') = \frac{1}{\sqrt{T}} \int_0^T (-)\langle K^a_{\theta'}, dX_t - H_{\theta'}(X_t)dt \rangle \qquad a = 1, \ldots, p \tag{69}$$

Crucially, this is not just a change of notation aimed at hiding away the difference between H_θ and D_θ. The new Itô differential is well defined in its own right. For instance it can be approximated numerically using Geodesic interpolation as defined by Emery [18] or Darling [26].

Roughly, geodesic interpolation is a map which associates to any two points $x, y \in M$ which are close enough to each other the vector $K = I(x,y) \in T_xM$ such that the geodesic γ with $\gamma(0) = x$ and $\dot{\gamma}(0) = K$ verifies $\gamma(1) = y$, (the dot denotes the velocity vector). If the word "geodesic" means a geodesic of the Le Jan-Watanabe connection $\bar{\nabla}$, then the following can be used for numerical approximation

$$w^a_T(\theta') = \frac{1}{\sqrt{T}} P_\theta - \lim_{\delta \to 0} \sum_{k\delta < T} \left\langle K^a_{\theta'}\left(X_{(k-1)\delta}\right), I\left(X_{(k-1)\delta}, X_{k\delta}\right) - H_{\theta'}\left(X_{(k-1)\delta}\right) \times \delta \right\rangle \tag{70}$$

where $\delta > 0$ is a step size. In particular, $K^a_{\theta'} = \partial_a H_{\theta'}$ can be used to compute numerically the maximum likelihood estimate θ^*_T. A more intuitive notation can be used to in writing (70). Let $\delta X_{k-1} = I(X_{(k-1)\delta}, X_{k\delta})$ and $X_{k-1} = X_{(k-1)\delta}$. Then, (70) becomes

$$w^a_T(\theta') = \frac{1}{\sqrt{T}} P_\theta - \lim_{\delta \to 0} \sum_{k\delta < T} \langle K^a_{\theta'}\left(X_{k-1}\right), \delta X_{k-1} - H_{\theta'}\left(X_{k-1}\right) \times \delta \rangle \tag{71}$$

Now, the method of estimation based on searching for ρ^*_T such that $w_T(\rho^*_T) = 0$ simply expresses the fact that $\delta X_{k-1} - H_{\theta'}\left(X_{k-1}\right) \times \delta$, which converges to zero as $\delta \to 0$, has an asymptotically normal distribution for small δ. This normal distribution has zero mean. Moreover, in the in the orthonormal frame $(E^i; i = 1, \ldots, d)$, its covariance matrix is δI_d where I_d is the $d \times d$ identity matrix.

7. Diffusions in Lie groups and symmetric spaces

In this section, two closely related examples are considered. First, in Paragraph 7.1, the manifold M is taken to be a Lie group and the observation process X a right invariant diffusion, on this Lie group. Second, in Paragraph 7.2, M is a symmetric space, under the action of a connected semisimple Lie group G, and X is a diffusion in M induced by this group action.

In either case, the aim will be to write down the likelihood equation (31) in a concrete form, in terms of the Lie group or symmetric space structure. This will require discussing conditions under which this equation is valid. Such conditions include, at least, hypotheses **(H1-H3)**.

Invariant diffusions on Lie groups, and more generally Lévy processes on Lie groups, have generated much recent attention. A thorough account can be found in Liao's book [27], which also addresses Brownian motion in symmetric spaces.

A major reference on the differential geometry of Lie groups and symmetric spaces is Helgason's monograph [28].

It is here useful to make some remarks, placing the current section in the general context of the paper.

The claim was made in the introduction that the paper studies estimation problems involving a general diffusion on a general manifold. In particular, this means there is no *a priori* relation between the diffusion process X and any additional structure on the manifold M.

The present section studies precisely the special case where X is compatible with the Lie group or symmetric space structure of the M. Naturally, this leads to certain simplifications which, in effect, make up much of the following. Using the general framework of this paper, it is possible to go beyond this special case. For instance, if M is a Lie group, one may be interested in a diffusion X with stationary density of the form (20), (roughly, this density represents a Gibbs distribution), which is not compatible with the Lie group structure, except in trivial cases. Then, the Lie group structure of M plays a limited role, while the properties of the diffusion X come to the forefront.

7.1. Invariant diffusion in a Lie group

In this paragraph, it is assumed the manifold M is a Lie group. Sticking to convention, M is then denoted G. It is clearly interesting to consider the case where the observation process X is compatible with the Lie group structure of G.

Here, X is taken to be a right invariant diffusion, parameterised by $\theta \in \Theta$. The aim will be to describe the maximum likelihood equation (31), of Paragraph 3.2. First, it is suitable to discuss hypotheses **(H1-H3)**.

Hypothesis **(H1)** refers to equation (1). In the present context, this equation is given as follows.

Let e be the identity element of G. Also, let $\mathfrak{g}$ be the Lie algebra of G, identified as the tangent space T_eG. Fix a basis $(\sigma_r; r = 1, \ldots, d)$ of $\mathfrak{g}$ and denote $(V_r; r = 1, \ldots, d)$ the corresponding family, (in fact, basis), of right invariant vector fields [28].

For simplicity, assume $\Theta = \mathbb{R}^d$. In other words, the dimension of the parameter space is the same as the number of vector fields V_r, (this is an identifiability assumption, similar to hypothesis **(H4)**).

With this in mind, consider

$$H_\theta(g) = \sum_{r=1}^{d} \theta^r V_r(g) \qquad (\theta, g) \in \Theta \times G \tag{72}$$

Equation (1) can then be transcribed

$$dY_t = \sum_{r=1}^{d} V_r(Y_t) \circ dy_t^r \qquad dy_t^r = \theta^r dt + dB_t^r \tag{73}$$

Since it only involves right invariant vector fields, equation (73) is said to be right invariant. Precisely, for any $h \in G$, if Y solves this equation then so does Y^h, where $Y_t^h = Y_t h$ is the product of Y_t and h in the group G.

Now, it can be seen hypothesis **(H1)** holds without any additional assumptions. In particular, there is no need to impose the conditions in Appendix A.

This follows by the argument of McKean, in Section 4.7 of [29]. Precisely, by Ado's theorem, it is always possible to assume G is a matrix Lie group. Then, (73) is shown to be a linear matrix stochastic differential equation, so that it has a unique strong solution Y defined for all $t \geq 0$, independently of its initial condition.

Consider hypothesis **(H2)**. The metric $\langle \cdot, \cdot \rangle$ of (5) is right invariant and completely defined by the statement that $(V_r(g); r = 1, \ldots, d)$ is an orthonormal basis of T_gG for all $g \in G$.

The corresponding volume measure v is a right Haar measure of G. Intuitively, since equation (73) is right invariant, the probability measure μ^*_θ must also be right invariant — Recall that μ^*_θ appears in (15) of Paragraph 2.3. It follows that μ^*_θ, when it exists, is a constant multiple of the Haar measure v. A rigorous form of this reasoning can be found in [27].

To conclude, when G is compact, therefore of finite volume, hypothesis **(H2)** is verified, with p_θ a normalising constant independent of θ. When G is not compact, hypothesis **(H2)** is not verified.

Finally, hypothesis **(H3)** is trivially verified. Since $H_\theta(g)$ is defined by (72) and the vectors $V_r(g)$ are orthonormal, it follows for $\theta', \theta \in \Theta$ that $\|H_{\theta'}(g) - H_\theta(g)\| = |\theta' - \theta|$, which is uniformly bounded in g. Recall here $|\cdot|$ is the Euclidean norm on Θ.

When G is compact, hypotheses **(H1-H3)** are verified. The likelihood equation (31) then reads

$$\int_0^T \left\langle V_r, dX_t - \sum_{u=1}^d \theta^u_T V_u(X_t)dt \right\rangle = 0 \qquad r = 1, \ldots, d \tag{74}$$

where $(\theta^u_T; u = 1, \ldots, d)$ denote the components of the maximum likelihood estimate θ^*_T, (which is a random element of $\mathbb{R}^d$).

To see (74) is indeed the likelihood equation, recall this equation is obtained by setting equal to zero the score function $\partial\ell_T$ given by (30).

It has been stated that, in the present context, p_θ is independent of θ, so that the first term on the right hand side of (30) becomes identically zero. The second term on the right hand side of (30) should be written down according to the definition of D_θ, from Theorem 1. Recall the expression of the Levi-Civita connection

$$\nabla_{V_r} V_u = \frac{1}{2}[V_r, V_u] \tag{75}$$

where $[\cdot, \cdot]$ denotes the Lie bracket of two vector fields. In particular, $[V_r, V_r] = 0$ and therefore $D_\theta = H_\theta$. Replacing in (30), the score function $\partial\ell_T$ follows from a simple calculation and (74) can be obtained immediately.

To apply equation (74) in practice, recall the vector fields $(V_r; r = 1, \ldots, d)$ define an orthonormal basis $(V_r(g); r = 1, \ldots, d)$ in each tangent space T_gG. This implies the solution of (74) is given by

$$\theta^r_T = \frac{1}{T} \int_0^T \langle V_r, dX_t \rangle \tag{76}$$

For concreteness, assume now G is a matrix Lie group. Formula (76), for the maximum likelihood estimate, admits the following simplification. Formally, recalling the metric $\langle \cdot, \cdot \rangle$ is right invariant, it is possible to write

$$\langle V_r(X_t), dX_t \rangle = \langle V_r(X_t)X_t^{-1}, dX_t X_t^{-1} \rangle$$

Here, X_t is a random matrix and X_t^{-1} its inverse matrix, both of them with their values in the group G. Since the vector fields V_r are right invariant, $V_r(g)g^{-1} = V_r(e) = \sigma_r$

for $g \in G$. Replacing in (76), it follows

$$\theta^r_T = \frac{1}{T}\int_0^T \langle \sigma_r, dX_t X_t^{-1}\rangle \tag{77}$$

The correct interpretation of this formula is that $dX_t X_t^{-1}$ is a matrix of Itô stochastic differentials which is formed, (according to the standard rule for matrix product), as the product of the matrices dX_t and X_t^{-1}. A rigorous justification of the equivalence between (76) and (77) is given in [30].

It is possible to express (77) using only matrix operations. That is, without any reference to differential geometry on the Lie group G. Recall that $(\sigma_r; r = 1, \dots, d)$ are orthonormal. This implies

$$\hat{\theta}^*_T = \frac{1}{T}\int_0^T dX_t X_t^{-1} \tag{78}$$

where $\hat{\theta}^*_T = \sum_{r=1}^d \theta^r_T \sigma_r$. Since $(\sigma_r; r = 1, \dots, d)$ are linearly independent, $\hat{\theta}^*_T$ and θ^*_T determine each other uniquely. Formula (78) gives an expression of the maximum likelihood estimate θ^*_T, which is immediately applicable whenever G is a matrix Lie group. For particular matrix Lie groups, this formula can be simplified even further. For example, if G is a group of orthogonal or unitary matrices, the matrix inverse under the integral can be replaced by a transpose or Hermitian transpose.

In the above discussion, the condition that the Lie group G is compact was imposed in order to ensure that hypothesis **(H2)** holds. It did not play any role in the discussion of the likelihood equation (74) and its solution.

Even when G is not compact, equation (74) is still well defined and its (unique) solution given by (76), or equivalently (78). However, in this case of non compact G, equation (74) cannot be termed a likelihood equation, without some substitute assumption as to the distribution of X_0 being made.

Without assuming G is compact, it is possible to prove that θ^*_T, defined by (76), is consistent and normal, and therefore asymptotically normal. These are the same conclusions as in Propositions 2 and 3 of Section 4. Note that, for $\theta \in \Theta$, by (76),

$$\theta^r_T - \theta^r = \frac{1}{T}\int_0^T \left\langle V_r, dX_t - \sum_{u=1}^d \theta^u V_u(X_t)dt \right\rangle$$

This is verified by direct calculation, using the fact that the vector fields V_r are orthonormal in each tangent space of G. Using (11), (12) and Lévy's characterisation of Brownian motion, (compare to the proof of Theorem 1), it follows that

$$\mathcal{L}_\theta\{\sqrt{T}(\theta^*_T - \theta)\} = N(I_d) \tag{79}$$

where I_d is the $d \times d$ identity matrix. This shows θ^*_T is indeed consistent and normal.

The matrix expression for the maximum likelihood estimate, given by (78), is essentially the same as suggested by Willsky and Lo [8, 9], who directly considered the special case where G is a matrix Lie group.

7.2. Induced diffusion in a symmetric space

In this paragraph, the manifold M is taken to be a simply connected symmetric space and the observation process X an induced diffusion in M, parameterised by $\theta \in \Theta$, (the term "induced diffusion" is used in [27]). The aim will be to give a concrete expression of the likelihood equation (31).

To begin, assume a connected semisimple Lie group G acts on M transitively so $M = G/K$ where K is the isotropy group of some point $o \in M$. The Lie algebras of G and K are denoted $\mathfrak{g}$ and $\mathfrak{k}$, respectively. Recall the following basic property [28]. There exists a scalar product $(\cdot,\cdot)$ on $\mathfrak{g}$, which is $\mathrm{Ad}(K)$-invariant. Accordingly, if $\mathfrak{m}$ denotes the orthogonal complement of $\mathfrak{k}$, then $\mathfrak{m}$ is also $\mathrm{Ad}(K)$-invariant.

As a special case of this setting, spaces of constant curvature can be obtained, (spherical, Euclidean and hyperbolic spaces). Precisely [31], M has constant curvature if and only if all orthogonal transformation O of $\mathfrak{m}$ is of the form $O = \mathrm{Ad}(k)$, $k \in K$.

Induced diffusions in spaces of constant curvature arise in many engineering problems. Consider, for example, the propagation of electromagnetic waves in random media. In [32], propagation in random lossless optical fibres was modeled using induced diffusions in the unit sphere S^2, considered with the action of the rotation group $SO(3)$. To model propagation in lossy optical fibres, it is necessary to consider induced diffusions in the light cone, considered with the action of the Lorentz group $SO(1,3)$. Besides propagation in random optical fibres, propagation in random transmission lines can be modeled using induced diffusions in the Poincaré unit disc, which is a model of the basic hyperbolic space, with the action of $SL(2,\mathbb{R})$ — See [33]. Note that randomness in optical fibres, transmission lines, or other propagation media, is a physical effect due to the presence of inhomogeneities which destroy the coherency of electromagnetic waves.

For simplicity, it will be assumed that M is embedded in some higher dimensional Euclidean space, where G acts as a matrix Lie group. The action of G on M is therefore denoted as a multiplication, $x \mapsto gx$, where $g \in G$ is a matrix and $x \in M$ a vector. To write down equation (1), let v be the dimension of G and $(\sigma_r; r = 1,\ldots,v)$ an orthonormal basis of $\mathfrak{g}$. For $x \in M$, let $V_r(x)$ be the vector in $T_x(M)$ given by

$$V_r(x) = \left.\frac{d}{dt}\right|_{t=0} \exp\left(t\sigma_r\right) x = \sigma_r x \tag{80}$$

where $\exp : \mathfrak{g} \to G$ is the matrix exponential.

Let $\Theta = \mathbb{R}^v$, as in the previous paragraph, and define

$$H_\theta(x) = \sum_{r=1}^{v} \theta^r V_r(x) \qquad (\theta, x) \in \Theta \times M \tag{81}$$

Now, equation (1) can be copied

$$dY_t = \sum_{r=1}^{v} V_r(Y_t) \circ dy_t^r \qquad dy_t^r = \theta^r dt + dB_t^r \tag{82}$$

This looks exactly like equation (73) from the previous paragraph. However, the vector fields V_r are here defined on the symmetric space M and not on the Lie group G. In particular, the number v of these vector fields is greater than the dimension d of M.

The relation between equations (73) and (82) is the following. Let R be a process with values in G which solves (73) with initial condition $R_0 = e$, (recall e is the identity element of G). If $x \in M$ then the process Y where $Y_t = R_t x$ solves (82) with initial condition x.

In other words, a diffusion X in M whose infinitesimal generator is given by (3), according to (80) and (81), can be induced by a right invariant diffusion with values in G. This justifies the name "induced diffusion" [27].

As in the previous paragraph, hypothesis **(H1)** holds without any additional assumptions. Replacing the expression of the vector fields V_r, from (81), shows equation (82) is a linear stochastic differential equation, so it has a unique strong solution Y, defined for all $t \geq 0$, given any initial condition.

In order to go on, it is necessary to discuss the metric $\langle \cdot, \cdot \rangle$ of (5) and the associated Levi-Civita connection ∇.

It turns out the metric $\langle \cdot, \cdot \rangle$ is induced by the scalar product $(\cdot, \cdot)$ on $\mathfrak{g}$. Define $V : \mathfrak{g} \times M \to TM$ to be the mapping, (this is a kind of repetition of (81)),

$$V(\sigma, x) = \left.\frac{d}{dt}\right|_{t=0} \exp(t\sigma)\, x = \sigma x \tag{83}$$

This will also be written $V(\sigma, x) = V_x \sigma$.

For fixed $x \in M$, this is a linear mapping $V_x : \mathfrak{g} \to T_x M$. Its kernel is denoted $\mathfrak{k}_x$. For example, $\mathfrak{k}_o = \mathfrak{k}$. The orthogonal complement of $\mathfrak{k}_x$ is denoted $\mathfrak{m}_x$, (this is the orthogonal complement with respect to $(\cdot, \cdot)$). Because G acts transitively on M, the mapping V_x is surjective for all $x \in M$. Its restriction to $\mathfrak{m}_x$ is an isomorphism between $\mathfrak{m}_x$ and $T_x M$. If $E \in T_x M$, its unique inverse image in $\mathfrak{m}_x$, under V_x, is denoted $V_x^{-1} E$ or $V^{-1}(E, x)$.

The metric $\langle \cdot, \cdot \rangle$ is given by

$$\langle E, K \rangle = \left(V_x^{-1} E, V_x^{-1} K\right) \qquad E, K \in T_x M \tag{84}$$

Moreover [28], it is invariant under the action of G on M. To show that (84) indeed verifies (5), note first that

$$V_x^{-1}\left(V_x \sigma\right) = \Pi_x(\sigma) \tag{85}$$

where Π_x denotes orthogonal projection of $\sigma \in \mathfrak{g}$ onto $\mathfrak{m}_x$. Using this identity, and the fact that $(\sigma_r; r = 1, \ldots, v)$ form an orthonormal basis of $\mathfrak{g}$, the right hand side of (84) can be written

$$\sum_{r=1}^{v} \left(V_x^{-1}(E), \Pi_x(\sigma_r)\right) \left(V_x^{-1}(K), \Pi_x(\sigma_r)\right) = \sum_{r=1}^{v} \langle E, V_r(x) \rangle \langle K, V_r(x) \rangle$$

which is (5). Note the fact that V_x^{-1} maps $\mathfrak{g}$ to $\mathfrak{m}_x$ allows σ_r to be replaced by $\Pi_x(\sigma_r)$.

Since $\langle \cdot, \cdot \rangle$ is invariant under the action of G on M, the corresponding volume measure v is also invariant under this action. Moreover, invariance under the action of G uniquely defines v. Using the corresponding result from the previous paragraph, it is now possible to show that hypothesis **(H2)** is verified if and only if G is compact. In this case, p_θ is a normalising constant independent of θ. Roughly, this follows by recalling, as stated above, that any solution Y of (82) is of the form $Y_t = R_t x$ where R is a solution of (73) and $x \in M$. For a rigorous formulation, see [27].

Hypothesis **(H3)**, is immediately verified. It can be shown that

$$\|H_{\theta'}(x) - H_\theta(x)\|^2 \leq |\theta' - \theta|^2$$

This is because, letting $\theta' - \theta = \lambda$, (using (81), (84) and (85)), the left hand side is

$$\sum_{r,u=1}^{v} \lambda^r \lambda^u \langle V_r(x), V_u(x) \rangle = \sum_{r,u=1}^{v} \lambda^r \lambda^u \left(\Pi_x(\sigma_r), \Pi_x(\sigma_u) \right) \leq \sum_{r,u=1}^{v} \lambda^r \lambda^u \left(\sigma_r, \sigma_u \right) = |\lambda|^2$$

Thus, the hypothesis is verified, since the right hand side does not depend on x. In conclusion, hypotheses **(H1-H3)** are verified, as soon as G is compact.

Assuming G is compact, the likelihood equation (31), in the present case, takes on a form very similar to (74). Precisely, (31) can be written

$$\int_0^T \left\langle V_r, dX_t - \sum_{u=1}^{v} \theta^u_T V_u(X_t) dt \right\rangle = 0 \qquad r = 1, \ldots, v \tag{86}$$

Here, the notation θ^u_T is the same as in equation (74) of the previous paragraph. That this is indeed equation (31) should be verified by expressing the score function $\partial \ell_T$ of (30). The first term on the right hand side of (30) is identically zero, because p_θ does not depend on θ. For the second term, the definition of D_θ should be applied, (this was given in Theorem 1). This requires being able to evaluate $\sum_{r=1}^{v} \nabla_{V_r} V_r$ where ∇ is the Levi-Civita connection. In [34], (Theorem 1.4.8 on page 28), it is shown that $\sum_{r=1}^{v} \nabla_{V_r} V_r = 0$. Thus, (86) follows from (30).

Equation (86) is a linear equation for θ^*_T. It has a unique solution, for any value of T. Indeed, this equation reads, (recall the components of θ^*_T are the θ^u_T),

$$\sum_{u=1}^{v} \left(\int_0^T \langle V_r(X_t), V_u(X_t) \rangle dt \right) \theta^u_T = \int_0^T \langle V_r, dX_t \rangle \qquad r = 1, \ldots, v \tag{87}$$

Moreover, the $v \times v$ matrix with elements $\langle V_r(X_t), V_u(X_t) \rangle$ is strictly positive definite.

Since equation (86), or (87), is well defined, its solution only required a matrix inversion. Evaluation of the matrix on the left hand side only involves an ordinary integral. The stochastic integral on the right hand side can be approximated using geodesic interpolation as indicated in Section 6.

8. Conclusion

The method of maximum likelihood estimation studied in this paper has several advantages. It has a clear interpretation and leads, mostly, to analytically tractable expressions. Its numerical implementation, as briefly discussed in Section 6, is often feasible but becomes completely unreliable when the sampling frequency, (inverse of the step size δ appearing in (70) and (71)), cannot be made large enough. This is not surprising as the method was developed from the start with the assumption of continuous time observation.

Even when this assumption holds, it may be useful to consider other estimation methods, such as the generalised method of moments [35]. In particular, the generalised method of moments is quite straightforward to apply in the case of a reversible diffusion, (given by (20) of Paragraph 2.3). Unfortunately, due to lack of space, this could not be detailed here.

In the case where the assumption of continuous time observation cannot be considered to hold, the parameter estimation problem becomes significantly harder. Developing general estimation methods with guaranteed performance, for this case, is an interesting topic for future research. Of course, it is always possible to pursue a brute force Monte Carlo method, in order to simulate the likelihood function of discrete time observation. In the Euclidean case, this was done in [36]. Analytically, it is very difficult to know the exact form or even the properties of this function.

Other approaches may lead to tractable methods, which require less computational effort than direct Monte Carlo simulation. As an indication for future work, consider the two following methods.

Recall, from Section 6, that the maximum likelihood method for continuous time observation is based on the fact that the "geodesic increments" of the observation process are asymptotically normal, when the step size δ becomes small — See discussion after (71). When δ cannot be considered small, by analogy with the idea of [37] for the Euclidean case, one could try expanding the likelihood function in an Edgeworth series, in order to obtain an approximate likelihood equation.

The second method, particularly well suited for the case of a reversible diffusion, is to construct martingale estimating functions using the eigenfunctions of the infinitesimal generator. In the Euclidean case, this was developed in [38].

In addition to the assumption of continuous time observation, another fundamental assumption for this paper was that the observation process is an elliptic diffusion. Indeed, all geometric constructions used in the paper are based on the Riemannian metric, (given by (5)), defined by the ellipticity assumption.

An interesting direction, in which the current paper can be generalised, is dropping the assumption of ellipticity and replacing it by hypoellipticity. Roughly, this would not change much of the ergodic properties of the observation process, but would require a more advanced approach to the geometry of this process.

In conclusion, while the general problem of parameter estimation for diffusions in manifolds has received very little attention in the literature, the extensive development of parameter estimation methods for Euclidean diffusions, and of the tools stochastic differential geometry, makes it an interesting target for future research, where many new results can be found.

References

[1] Chirikjian, G. S. (2009). *Stochastic models, information theory and Lie groups, Volume 1*, Birkhäuser. Boston.

[2] Chirikjian, G. S. (2011). *Stochastic models, information theory and Lie groups, Volume 2*, Birkhäuser. Boston.

[3] Perrin, F. (1928). Étude mathématique du mouvement Brownien de rotation. *Annales Scientifiques de l'É.N.S.* 3^{e} *Série* **45,** 1–51.

[4] Yosida, K. (1949). Integration of Fokker-Planck's equation in a compact Riemannian space. *Arkiv fur Matematik* **1,** 71–75.

[5] Itô, K. (1950). Brownian motions in a Lie group. *Proc. Japan. Acad.* **26,** 4–10.

[6] McKean, H. P. (1960). Brownian motions on the 3-dimensional rotation group. *Memoirs of the College of Science, University of Kyoto, Series A* **33,** 25–38.

[7] Ikeda, N. and Watanabe, Sh. (1981). *Stochastic differential equations and diffusion processes*, North-Holland Publishing Company.

[8] Willsky, A. S. (1973). *Dynamical systems defined in groups: Structural properties and estimation*, Ph.D. dissertation, Massachusetts Institute of Technology.

[9] Lo, J. T. H., *Signal detection for bilinear systems*, Information Sciences, 9 (1975), pp. 249–278.

[10] Said, S. and Manton, J. H. (2012). Extrinsic mean of Brownian distributions on compact Lie groups. *IEEE Trans. Inf Theory.* **58,** 3521–3535.

[11] Wolfe, K. C. and Chirikjian, G. S., *Signal detection on Euclidean groups: Applications to DNA bends, robot localization and optical communication*, *IEEE J. Selected Topics in Signal Processing*, 7 (2013), pp. 708–719.

[12] Ng, S. K. and Caines, P. E. and Chen, H. F. (1984). Parameter estimation for observed diffusions in manifolds. *IMA J. Math. Control Inf.* **1,** 129–140.

[13] Duncan, T. E. (1977). Some filtering results in Riemann manifolds. *Information and Control* **35,** 182–195.

[14] Ng, S. K. and Caines, P. E. (1985). Nonlinear filtering in Riemannian manifolds. *IMA J. Math. Control Inf.* **2,** 25–36.

[15] Pontier, M. and Szpirglas, J. (1985). Filtrage non linéaire avec observation sur une variété. *Stochastics* **15,** 121–148.

[16] Said, S. and Manton, J. H. (2013). Filtering with observation in a manifold: Reduction to a classical filtering problem. *SIAM J. Control Optim.* **51,** 767–783.

[17] Kutoyants, Y. (2004). *Statistical inference for ergodic diffusion processes*, Springer-Verlag. London.

[18] Emery, M. (1989). *Stochastic calculus in manifolds*, Springer-Verlag.

[19] Hsu, E. P. (2002). *Stochastic analysis on manifolds*, American Mathematical Society.

[20] Heyde, C. C. (1997). *Quasi-likelihood and its applications: A general approach to optimal parameter estimation*, Springer. New York.

[21] Deuschel, J. D. and Stroock, D. W. (1989). *Large deviation*, Academic Press.

[22] Bain, A. and Crisan, D. (2010). *Fundamentals of stochastic filtering*, Springer Science.

[23] Karandikar, R. L. (1983). Interchanging the order of stochastic integration and ordinary differentiation. *Sankhya A* **45,** 120–124.

[24] Kallenberg, O. (2002). *Foundations of modern probability*, 2nd Edition, Springer-Verlag.

[25] Rao, C. R. (1973). *Linear statistical inference and its applications*, John Wiley & Sons.

[26] Darling, R. W. R. (1982). *Martingales on manifolds and geometric Itô calculus*, PhD Thesis, The University of Warwick.

[27] Liao, M. (2004). *Lévy processes in Lie groups*, Cambridge University Press.

[28] Helgason, S. (1962). *Differential geometry and symmetric spaces*, Academic Press.

[29] McKean, H. P. (1969). *Stochastic integrals*, Academic Press.

[30] Hakim-Dowek, M. and Lépingle, D., *L'exponentielle stochastique des groupes de Lie*, Séminaire de Probabilités (Strasbourg), 20 (1986), pp. 352–374.

[31] Wolf, J. A. (2010). *Spaces of constant curvature*, 6th Edition, AMS Chelsea Publishing.

[32] Said, S. and Le Bihan, N. (2008). Higher order statistics of Stokes parameters in a random birefringent medium. *Waves Random Complex Media* **18,** 275–292.

[33] Terras, A. (1984). Noneuclidean Harmonic analysis, the central limit theorem and long transmission lines with random inhomogeneities. *J. Multivariate Anal.* **115,** 261–276.

[34] Elworthy, K. D. and Le Jan Y. and Li X. M. (1999). *On the geometry of diffusion operators and stochastic flows*, Springer-Verlag.

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

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

[37] Aït-Sahalia, Y. (2002). Maximum likelihood estimation for discretely sampled diffusions: A closed form approximation approach. *Econometrica* **70,** 223–262.

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

Appendix A. Stochastic completeness and hypothesis (H1)

The current appendix is aimed at giving sufficient conditions which guarantee hypothesis **(H1)** holds. These conditions, given in Proposition 6 below, mainly involve the Riemannian geometry of the manifold M with metric (5). In many cases, this means they are relatively easy to check and thus of practical use.

Recall hypothesis **(H1)** states existence and uniqueness of weak solutions of equation (1), for each value of the parameter $\theta \in \Theta$.

Precisely, existence of a weak solution means it is possible to construct a probability space, on which a Brownian motion B and a process Y^x_θ are defined that together satisfy (1). Here, uniqueness of weak solutions is taken to mean that the distribution P^x_θ of Y^x_θ is uniquely determined by x and θ.

Note, in particular, that hypothesis **(H1)** requires the solution Y^x_θ of (1) is defined for all $t \geq 0$. That is, Y^x_θ does not explode. Hypothesis **(H1)** is always verified when M is compact. Compactness of M even guarantees existene and pathwise uniqueness of strong solutions [7].

The main proposition, Proposition 6 below, does not require M to be compact. Rather, it is invokes the geometric notion of stochastic completeness. The Riemannian manifold M, with metric (5), is called stochastically complete if a Brownian motion started at any point $x \in M$ does not explode.

Precisely, M is stochastically complete if the following equation has a unique weak solution Y^x defined for all $t \geq 0$,

$$dY_t = -\frac{1}{2}\sum_{r=1}^{v} \nabla_{V_r} V_r(Y_t)dt + \sum_{r=1}^{v} V_r(Y_t) \circ dB^r_t \qquad Y_0 = x \tag{88}$$

Here, $x \in M$ is any deterministic initial condition. Applying the classical Itô formula, it is straightforward that for any smooth function f on M,

$$df(Y_t) = \frac{1}{2}\Delta f(Y_t)dt + \sum_{r=1}^{v} V_r f(Y_t)dB_t^r \tag{89}$$

whenever Y is a weak solution of (88). Thus, in this case, Y solves the martingale problem associated to the Laplacian Δ of M — Recall the definition of Δ from (6). This is the usual definition of Brownian motion on a Riemannian manifold.

There are many simple sufficient conditions for stochastic completeness. For example, all Riemannian manifolds whose Ricci curvature is bounded below are stochastically complete. More generally, all Riemannian manifolds with polynomial volume growth are stochastically complete. Such conditions can be checked using classical results in Riemannian geometry [19].

It is worth mentioning that stochastic completeness is a not implied by geodesic completeness, (Proposition 4.2.6 on page 111 of [19]).

The idea of Proposition 6 is that when M is stochastically complete, a weak solution of (1) can be obtained from a weak solution of (88) by a Change of measure, using Girsanov's theorem.

Proposition 6. *Assume M is stochastically complete. If, for $\theta \in \Theta$, $\sup_{x\in M} \|D_\theta(x)\| < +\infty$, then hypothesis **(H1)** holds.*

Proof. Since M is stochastically complete, there exists some probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$, on which a Brownian motion B and a process Y^x, (for each $x \in M$), are defined such that (88) is satisfied. Moreover, Y^x is defined with values in M for all $t \geq 0$.

For any smooth function f on M, the process Y^x verifies (89). Written out, using (6), this becomes

$$df(Y_t^x) = \frac{1}{2}\sum_{r=1}^{v}\{V_r^2 f(Y_t^x) - \nabla_{V_r}V_r f(Y_t^x)\}dt + \sum_{r=1}^{v} V_r f(Y_t^x)dB_t^r \qquad t \geq 0 \tag{90}$$

Since $\|D_\theta\|$ is bounded, the following process L is a $\tilde{P}$-martingale, (with respect the augmented filtration generated by Y^x and B),

$$L_t = \exp\left(\int_0^t \sum_{r=1}^{v}\langle D_\theta(Y_t^x), V_r(Y_t^x)\rangle dB_t^r - \frac{1}{2}\int_0^t \|D_\theta(Y_t^x)\|^2 dt\right) \tag{91}$$

Then, there exists a probability measure $\tilde{P}_\theta$ on $\tilde{\mathcal{F}}$ such that

$$\left.\frac{d\tilde{P}_\theta}{d\tilde{P}}\right|_t = L_t \qquad t \geq 0 \tag{92}$$

By Girsanov's theorem $B(\theta)$, defined by the following formula, is a $\tilde{P}_\theta$-Brownian motion

$$dB_t^r(\theta) = dB_t^r - \langle D_\theta(Y_t^x), V_r(Y_t^x)\rangle dt \tag{93}$$

From (90), using the fact that $D_\theta f = \sum_{r=1}^{v}\langle D_\theta, V_r\rangle V_r f$,

$$df(Y_t^x) = \{H_\theta f(Y_t^x) + \frac{1}{2}\sum_{r=1}^{v} V_r^2 f(Y_t^x)\}dt + \sum_{r=1}^{v} V_r f(Y_t^x)dB_t^r(\theta) \qquad t \geq 0$$

which is the same as

$$df(Y^x_t) = H_\theta f(Y^x_t)dt + \sum_{r=1}^{v} V_r f(Y^x_t) \circ dB^r_t(\theta) \qquad t \geq 0$$

Since this holds for any smooth function f, it follows that Y^x and $B(\theta)$ together satisfy the equation (1) with respect to $\tilde{P}_\theta$. This shows the existence of a weak solution defined for all $t \geq 0$.

Uniqueness of the distribution P^x_θ of Y^x under $\tilde{P}_\theta$ follows from uniqueness of the distribution of the same process Y^x under $\tilde{P}$. This is because $L_t > 0$, so $\tilde{P}_\theta$ and $\tilde{P}$ are equivalent. Note that uniqueness of the distribution of Y^x under $\tilde{P}$ is due to the uniqueness of the fundamental solution of the heat equation on M.