entropy

MDPI

Article

Fast, Asymptotically Efficient, Recursive Estimation in a Riemannian Manifold

Jialun Zhou [1,*,†] **and Salem Said** [2,†]

1 Department of Science and Technology, University of Bordeaux, 33076 Bordeaux, France
2 IMS Laboratory, University of Bordeaux, 33076 Bordeaux, France; salem.said@u-bordeaux.fr
* Correspondence: jialun.zhou@u-bordeaux.fr
† Current address: 351 Cours de la libération, 33405 Talence cedex, France.

Received: 3 September 2019; Accepted: 17 October 2019; Published: 21 October 2019

Abstract: Stochastic optimisation in Riemannian manifolds, especially the Riemannian stochastic gradient method, has attracted much recent attention. The present work applies stochastic optimisation to the task of recursive estimation of a statistical parameter which belongs to a Riemannian manifold. Roughly, this task amounts to stochastic minimisation of a statistical divergence function. The following problem is considered: how to obtain fast, asymptotically efficient, recursive estimates, using a Riemannian stochastic optimisation algorithm with decreasing step sizes. In solving this problem, several original results are introduced. First, without any convexity assumptions on the divergence function, we proved that, with an adequate choice of step sizes, the algorithm computes recursive estimates which achieve a fast non-asymptotic rate of convergence. Second, the asymptotic normality of these recursive estimates is proved by employing a novel linearisation technique. Third, it is proved that, when the Fisher information metric is used to guide the algorithm, these recursive estimates achieve an optimal asymptotic rate of convergence, in the sense that they become asymptotically efficient. These results, while relatively familiar in the Euclidean context, are here formulated and proved for the first time in the Riemannian context. In addition, they are illustrated with a numerical application to the recursive estimation of elliptically contoured distributions.

Keywords: Riemannian stochastic gradient; Fisher information metric; recursive estimation; asymptotic efficiency; elliptically contoured distributions

1. Introduction

Over the last five years, the data science community has devoted significant attention to stochastic optimisation in Riemannian manifolds. This was impulsed by Bonnabel, who proved the convergence of the Riemannian stochastic gradient method [1]. Later on [2], the rate of convergence of this method was studied in detail and under various convexity assumptions on the cost function. More recently, asymptotic efficiency of the averaged Riemannian stochastic gradient method was proved in [3]. Previously, for the specific problem of computing Riemannian means, several results on the convergence and asymptotic normality of Riemannian stochastic optimisation methods had been obtained [4,5]. The framework of stochastic optimisation in Riemannian manifolds is far-reaching, and encompasses applications to principal component analysis, dictionary learning, and tensor decomposition, to give only a few examples [6–8].

The present work moves in a different direction, focusing on recursive estimation in Riemannian manifolds. While recursive estimation is a special case of stochastic optimisation, it has its own geometric structure, given by the Fisher information metric. Here, several original results will be introduced, which show how this geometric structure can be exploited, to design Riemannian stochastic

optimisation algorithms which compute fast, asymptotically efficient, recursive estimates, of a statistical parameter which belongs to a Riemannian manifold. For the first time in the literature, these results extend, from the Euclidean context to the Riemannian context, the classical results of [9,10].

The mathematical problem, considered in the present work, is formulated in Section 2. This involves a parameterised statistical model P of probability distributions P_θ, where the statistical parameter θ belongs to a Riemannian manifold Θ. Given independent observations, with distribution P_{θ^*} for some $\theta^* \in \Theta$, the aim is to estimate the unknown parameter θ^*. In principle, this is done by minimising a statistical divergence function $D(\theta)$, which measures the dissimilarity between P_θ and P_{θ^*}. Taking advantage of the observations, there are two approaches to minimising $D(\theta)$: stochastic minimisation, which leads to recursive estimation, and empirical minimisation, which leads to classical techniques, such as maximum-likelihood estimation [11,12].

The original results, obtained in the present work, are stated in Section 3. In particular, these are Propositions 2, 4, and 5. Overall, these propositions show that recursive estimation, which requires less computational resources than maximum-likelihood estimation, can still achieve the same optimal performance, characterised by asymptotic efficiency [13,14].

To summarise these propositions, consider a sequence of recursive estimates θ_n, computed using a Riemannian stochastic optimisation algorithm with decreasing step sizes (n is the number of observations already processed by the algorithm). Informally, under assumptions which guarantee that θ^* is an attractive local minimum of $D(\theta)$, and that the algorithm is neither too noisy, nor too unstable, in the neighborhood of θ^*,

- Proposition 2 states that, with an adequate choice of step sizes, the θ_n achieve a fast non-asymptotic rate of convergence to θ^*. Precisely, the expectation of the squared Riemannian distance between θ_n and θ^* is $O\left(n^{-1}\right)$. This is called a fast rate, because it is the best achievable, for any step sizes which are proportional to n^{-q} with $q \in (1/2, 1]$ [9,15]. Here, this rate is obtained without any convexity assumptions, for twice differentiable $D(\theta)$. It would still hold for non-differentiable, but strongly convex, $D(\theta)$ [2].
- Proposition 4 states that the distribution of the θ_n becomes asymptotically normal, centred at θ^*, when n grows increasingly large, and also characterises the corresponding asymptotic covariance matrix. This proposition is proved using a novel linearisation technique, which also plays a central role in [3].
- Proposition 5 states that, if the Riemannian manifold Θ is equipped with the Fisher information metric of the statistical model P, then Riemannian gradient descent with respect to this information metric, when used to minimise $D(\theta)$, computes recursive estimates θ_n which are asymptotically efficient, achieving the optimal asymptotic rate of convergence, given by the Cramér-Rao lower bound. This is illustrated, with a numerical application to the recursive estimation of elliptically contoured distributions, in Section 4.

The end result of Proposition 5 is asymptotic efficiency, achieved using the Fisher information metric. In [3], an alternative route to asymptotic efficiency is proposed, using the averaged Riemannian stochastic gradient method. This method does not require any prior knowledge of the Fisher information metric, but has an additional computational cost, which comes from computing on-line Riemannian averages.

The proofs of Propositions 2, 4, and 5, are detailed in Section 6, and Appendices A and B. Necessary background, about the Fisher information metric (in short, this will be called the information metric), is recalled in Appendix C. Before going on, the reader should note that the summation convention of differential geometry is used throughout the following, when working in local coordinates.

2. Problem Statement

Let $P = (P, \Theta, X)$ be a statistical model, with parameter space Θ and sample space X. To each $\theta \in \Theta$, the model P associates a probability distribution P_θ on X. Here, Θ is a C^r Riemannian manifold

with $r > 3$, and X is any measurable space. The Riemannian metric of Θ will be denoted $\langle \cdot, \cdot \rangle$, with its Riemannian distance $d(\cdot, \cdot)$. In general, the metric $\langle \cdot, \cdot \rangle$ is not the information metric of the model P.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, and $(x_n ; n = 1, 2, \dots)$ be i.i.d. random variables on Ω, with values in X. While the distribution of x_n is unknown, it is assumed to belong to the model P. That is, $\mathbb{P} \circ x_n^{-1} = P_{\theta^*}$ for some $\theta^* \in \Theta$, to be called the true parameter.

Consider the following problem : how to obtain fast, asymptotically efficient, recursive estimates θ_n of the true parameter θ^*, based on observations of the random variables x_n? The present work proposes to solve this problem through a detailed study of the decreasing-step-size algorithm, which computes, similar to [1]

$$\theta_{n+1} = \mathrm{Exp}_{\theta_n}(\gamma_{n+1} u(\theta_n, x_{n+1})) \qquad n = 0, 1, \dots \tag{1a}$$

starting from an initial guess θ_0 .

This algorithm has three ingredients. First, Exp denotes the Riemannian exponential map of the metric $\langle \cdot, \cdot \rangle$ of Θ [16]. Second, the step sizes γ_n are strictly positive, decreasing, and verify the usual conditions for stochastic approximation [10,17]

$$\sum \gamma_n = \infty \qquad \sum \gamma_n^2 < \infty \tag{1b}$$

Third, $u(\theta, x)$ is a continuous vector field on Θ for each $x \in X$, which generalises the classical concept of score statistic [13,18]. It will become clear, from the results given in Section 3, that the solution of the above-stated problem depends on the choice of each one of these three ingredients.

A priori knowledge about the model P is injected into Algorithm (1a) using a divergence function $D(\theta) = D(P_{\theta^*}, P_\theta)$ (note that θ^* is unknown, though). As defined in [19], this is a positive function, equal to zero if and only if $P_\theta = P_{\theta^*}$, and with positive definite Hessian at $\theta = \theta^*$. Since one expects that minimising $D(\theta)$ will lead to estimating θ^*, it is natural to require that

$$E_{\theta^*} u(\theta, x) = -\nabla D(\theta) \tag{1c}$$

In other words, that $u(\theta, x)$ is an unbiased estimator of minus the Riemannian gradient of $D(\theta)$. With $u(\theta, x)$ given by (1c), Algorithm (1a) is a Riemannian stochastic gradient descent, of the form considered in [1–3]. However, as explained in Remark 2, (1c) may be replaced by the weaker condition (9), which states that $D(\theta)$ is a Lyapunov function of Algorithm (1a), without affecting the results in Section 3. In this sense, Algorithm (1a) is more general than Riemannian stochastic gradient descent.

In practice, a suitable choice of $D(\theta)$ is often the Kullback-Leibler divergence [20],

$$D(\theta) = -E_{\theta^*} \log L(\theta) \qquad L(\theta) = \frac{dP_\theta}{dP_{\theta^*}} \tag{2a}$$

where P_θ is absolutely continuous with respect to P_{θ^*} with Radon-Nikodym derivative $L(\theta)$ (the likelihood function). Indeed, if $D(\theta)$ is chosen to be the Kullback-Leibler divergence, then (1c) is satisfied by

$$u(\theta, x) = \nabla \log L(\theta) \tag{2b}$$

which, in many practical situations, can be evaluated directly, without any knowledge of θ^* .

Before stating the main results, in the following Section 3, it may be helpful to recall some general background on recursive estimation [10]. For simplicity, let $D(\theta)$ be the Kullback-Leibler divergence (2a). The problem of estimating the true parameter θ^* is equivalent to the problem of finding a global minimum of $D(\theta)$. Of course, this problem cannot be tackled directly, since $D(\theta)$ cannot be computed without knowledge of θ^*. There exist two routes around this difficulty.

The first route is empirical minimisation, which replaces the expectation in (2a) with an empirical mean over observed data. Given the first n observations, instead of minimising $D(\theta)$, one minimises the empirical divergence $D_n(\theta)$,

$$D_n(\theta) \;=\; -\frac{1}{n}\sum_{m=1}^{n} \log L(\theta, x_m) \tag{3}$$

where L is the likelihood function of (2a). Now, given the minus sign ahead of the sum in (3), it is clear that minimising $D_n(\theta)$ amounts to maximising the sum of log-likelihoods. Thus, one is lead to the method of maximum-likelihood estimation.

It is well-known that maximum-likelihood estimation under general regularity conditions is asymptotically efficient [13]. Roughly, this means the maximum-likelihood estimator has the least possible asymptotic variance, equal to the inverse of the Fisher information. On the other hand, as the number n of observations grows, it can be especially difficult to deal with the empirical divergence $D_n(\theta)$ of Equation (3). In the process of searching for the minimum of $D_n(\theta)$, each evaluation of this function, or of its derivatives, will involve a massive number of operations, ultimately becoming unpractical.

Aiming to avoid this difficulty, the second route recursive estimation is based on observation-driven updates, following the general scheme of algorithm (1a). In this scheme, each new recursive estimate θ_{n+1} is computed using only the new observation x_{n+1}. Therefore, as the number of observations grows very large, the overall computational effort required by recursive estimation remains the same.

The main results in the following section show that recursive estimation can achieve the same asymptotic performance as maximum-likelihood estimation as the number n of observations grows large. As a word of caution, it should be mentioned that recursive estimation does not, in general, fare better than maximum-likelihood estimation for moderate values of the number n of observations, and models with a small number of parameters. However, it is a very desirable substitute to maximum-likelihood estimation for models with a large number of parameters, which typically require a very large number of observations in order to be estimated correctly.

3. Main Results

The motivation of the following Propositions 1 to 5 is to provide general conditions, which guarantee that Algorithm (1a) computes fast, asymptotically efficient, recursive estimates θ_n of the true parameter θ^*. In the statement of these propositions, it is implicitly assumed that conditions (1b) and (1c) are verified. Moreover, the following assumptions are considered.

- **(d1)** the divergence function $D(\theta)$ has an isolated stationary point at $\theta = \theta^*$, and Lipschitz gradient in a neighborhood of this point.
- **(d2)** this stationary point is moreover attractive: $D(\theta)$ is twice differentiable at $\theta = \theta^*$, with positive definite Hessian at this point.
- **(u1)** in a neighborhood of $\theta = \theta^*$, the function $V(\theta) = E_{\theta^*}\|u(\theta, x)\|^2$ is uniformly bounded.
- **(u2)** in a neighborhood of $\theta = \theta^*$, the function $R(\theta) = E_{\theta^*}\|u(\theta, x)\|^4$ is uniformly bounded.

For Assumption (d1), the definition of a Lipschitz vector field on a Riemannian manifold may be found in [21]. For Assumptions (u1) and (u2), $\|\cdot\|$ denotes the Riemannian norm.

Assumptions (u1) and (u2) are so-called moment control assumptions. They imply that the noise in Algorithm (1a) does not cause the iterates θ_n to diverge, and are also crucial to proving the asymptotic normality of these iterates.

Let Θ^* be a neighborhood of θ^* which verifies (d1), (u1), and (u2). Without loss of generality, it is assumed that Θ^* is compact and convex (see the definition of convexity in [16,22]). Then, Θ^* admits a

system of normal coordinates $(\theta^\alpha ; \alpha = 1, \dots, d)$ with origin at θ^*. With respect to these coordinates, denote the components of $u(\theta^*, x)$ by $u^\alpha(\theta^*)$ and let $\Sigma^* = (\Sigma^*_{\alpha\beta})$,

$$\Sigma^*_{\alpha\beta} \;=\; E_{\theta^*}\left[u^\alpha(\theta^*)\, u^\beta(\theta^*)\right] \tag{4a}$$

When (d2) is verified, denote the components of the Hessian of $D(\theta)$ at $\theta = \theta^*$ by $H = (H_{\alpha\beta})$,

$$H_{\alpha\beta} \;=\; \frac{\partial^2 D}{\partial\theta^\alpha \partial\theta^\beta}\bigg|_{\theta^\alpha = 0} \tag{4b}$$

Then, the matrix $H = (H_{\alpha\beta})$ is positive definite [23]. Denote by $\lambda > 0$ its smallest eigenvalue.

Propositions 1 to 5 require the condition that the recursive estimates θ_n are stable, which means that all the θ_n lie in Θ^*, almost surely. The need for this condition is discussed in Remark 3. Note that, if θ_n lies in Θ^*, then θ_n is determined by its normal coordinates θ_n^α.

Proposition 1 (consistency)**.** *assume (d1) and (u1) are verified, and the recursive estimates θ_n are stable. Then,* $\lim \theta_n = \theta^*$ *almost surely.*

Proposition 2 (mean-square rate)**.** *assume (d1), (d2) and (u1) are verified, the recursive estimates θ_n are stable, and $\gamma_n = \frac{a}{n}$ where $2\lambda a > 1$. Then*

$$\mathbb{E}\, d^2(\theta_n, \theta^*) \;=\; O\left(n^{-1}\right) \tag{5}$$

Proposition 3 (almost-sure rate)**.** *assume the conditions of Proposition 2 are verified. Then,*

$$d^2(\theta_n, \theta^*) \;=\; o(n^{-p}) \text{ for } p \in (0,1) \qquad \text{almost surely} \tag{6}$$

Proposition 4 (asymptotic normality)**.** *assume the conditions of Proposition 2, as well as (u2), are verified. Then, the distribution of the re-scaled coordinates $(n^{1/2}\theta_n^\alpha)$ converges to a centred d-variate normal distribution, with covariance matrix Σ given by Lyapunov's equation*

$$A\,\Sigma + \Sigma\, A \;=\; -a^2\,\Sigma^* \tag{7}$$

where $A = (A_{\alpha\beta})$ with $A_{\alpha\beta} = \frac{1}{2}\delta_{\alpha\beta} - aH_{\alpha\beta}$ (here, δ denotes Kronecker's delta).

Proposition 5 (asymptotic efficiency)**.** *assume the Riemannian metric $\langle\cdot,\cdot\rangle$ of Θ coincides with the information metric of the model P, and let $D(\theta)$ be the Kullback-Leibler divergence (2a). Further, assume (d1), (d2), (u1) and (u2) are verified, the recursive estimates θ_n are stable, and $\gamma_n = \frac{a}{n}$ where $2a > 1$. Then,*

(i) *the rates of convergence (5) and (6) hold true.*

(ii) *if $a = 1$, the distribution of the re-scaled coordinates $(n^{1/2}\theta_n^\alpha)$ converges to a centred d-variate normal distribution, with covariance matrix Σ^*.*

(iii) *if $a = 1$, and $u(\theta, x)$ is given by (2b), then Σ^* is the identity matrix, and the recursive estimates θ_n are asymptotically efficient.*

(iv) *the following rates of convergence also hold*

$$\mathbb{E}\, D(\theta_n) \;=\; O\left(n^{-1}\right) \tag{8a}$$

$$D(\theta_n) \;=\; o(n^{-p}) \text{ for } p \in (0,1) \qquad \text{almost surely} \tag{8b}$$

The following remarks are concerned with the scope of Assumptions (d1), (d2), (u1), and (u2), and with the applicability of Propositions 1 to 5.

Remark 1. *(d2), (u1) and (u2) do not depend on the Riemannian metric $\langle \cdot, \cdot \rangle$ of Θ. Precisely, if they are verified for one Riemannian metric on Θ, then they are verified for any Riemannian metric on Θ. Moreover, if the function $D(\theta)$ is C^2, then the same is true for (d1). In this case, Propositions 1 to 5 apply for any Riemannian metric on Θ, so that the choice of the metric $\langle \cdot, \cdot \rangle$ is a purely practical matter, to be decided according to applications.*

Remark 2. *the conclusion of Proposition 1 continues to hold, if (1c) is replaced by*

$$E_{\theta^*} \langle u(\theta, x), \nabla D(\theta) \rangle < 0 \text{ for } \theta \neq \theta^* \tag{9}$$

Then, it is even possible to preserve Propositions 2, 3, and 4, provided (d2) is replaced by the assumption that the mean vector field, $X(\theta) = E_{\theta^*} u(\theta, x)$, has an attractive stationary point at $\theta = \theta^*$. This generalisation of Propositions 1 to 4 can be achieved following essentially the same approach as laid out in Section 6. However, in the present work, it will not be carried out in detail.

Remark 3. *the condition that the recursive estimates θ_n are stable is standard in all prior work on stochastic optimisation in manifolds [1–3]. In practice, this condition can be enforced through replacing Algorithm (1a) by a so-called projected or truncated algorithm. This is identical to (1a), except that θ_n is projected back onto the neighborhood Θ^* of θ^*, whenever it falls outside of this neighborhood [10,17]. On the other hand, if the θ_n are not required to be stable, but (d1) and (u1) are replaced by global assumptions,*

(d1') *$D(\theta)$ has compact level sets and globally Lipschitz gradient.*

(u1') *$V(\theta) \leq C\,(1 + D(\theta))$ for some constant C and for all $\theta \in \Theta$.*

then, applying the same arguments as in the proof of Proposition 1, it follows that the θ_n converge to the set of stationary points of $D(\theta)$, almost surely.

Remark 4. *from (ii) and (iii) of Proposition 5, it follows that the distribution of $n\, d^2(\theta_n, \theta^*)$ converges to a χ^2-distribution with d degrees of freedom. This provides a practical means of confirming the asymptotic efficiency of the recursive estimates θ_n.*

4. Application: Estimation of ECD

Here, the conclusion of Proposition 5 is illustrated, by applying Algorithm (1a) to the estimation of elliptically contoured distributions (ECD) [24,25]. Precisely, in the notation of Section 2, let $\Theta = \mathcal{P}_m$ the space of $m \times m$ positive definite matrices, and $X = \mathbb{R}^m$. Moreover, let each P_θ have probability density function

$$p(x|\theta) \propto \exp\left[h\left(x^\dagger \theta^{-1} x\right) - \frac{1}{2}\log\det(\theta)\right] \qquad \theta \in \mathcal{P}_m, x \in \mathbb{R}^m \tag{10}$$

where $h : \mathbb{R} \to \mathbb{R}$ is fixed, has negative values, and is decreasing, and † denotes the transpose. Then, P_θ is called an ECD with scatter matrix θ. To begin, let $(x_n\,; n = 1, 2, \ldots)$ be i.i.d. random vectors in $\mathbb{R}^m$, with distribution P_{θ^*} given by (10), and consider the problem of estimating the true scatter matrix θ^*. The standard approach to this problem is based on maximum-likelihood estimation [25,26]. An original approach, based on recursive estimation, is now introduced using Algorithm (1a).

As in Proposition 5, the parameter space $\mathcal{P}_m$ will be equipped with the information metric of the statistical model P just described. In [27], it is proved that this information metric is an affine-invariant metric on $\mathcal{P}_m$. In other words, it is of the general form [28]

$$\langle u, u \rangle_\theta = I_1 \operatorname{tr}\left(\theta^{-1} u\right)^2 + I_2 \operatorname{tr}^2\left(\theta^{-1} u\right) \qquad u \in T_\theta \mathcal{P}_m \tag{11a}$$

parameterised by constants $I_1 > 0$ and $I_2 \geq 0$, where tr denotes the trace and tr^2 the squared trace, and where $T_\theta \mathcal{P}_m$ denotes the tangent space at θ to the manifold $\mathcal{P}_m$. Precisely [27], for the information metric of the model P,

$$I_1 = \frac{\varphi}{2m^2(m+2)} \qquad I_2 = \frac{\varphi}{m^2} - \frac{1}{4} \tag{11b}$$

where φ is a further constant, given by the expectation

$$\varphi \;=\; E_e \left[h'(x^\dagger x) \left(x^\dagger x \right) \right]^2 \tag{11c}$$

with $e \in \mathcal{P}_m$ the identity matrix, and h' the derivative of h. This expression of the information metric can now be used to specify Algorithm (1a).

First, since the information metric is affine-invariant, it is enough to recall that all affine-invariant metrics on $\mathcal{P}_m$ have the same Riemannian exponential map [25,29],

$$\text{Exp}_\theta(u) \;=\; \theta \exp\left(\theta^{-1} u \right) \tag{12a}$$

where exp denotes the matrix exponential. Second, as in (ii) of Proposition 5, choose the sequence of step sizes

$$\gamma_n = \frac{1}{n} \tag{12b}$$

Third, as in (iii) of Proposition 5, let $u(\theta, x)$ be the vector field on $\mathcal{P}_m$ given by (2b),

$$u(\theta, x) = \nabla^{(inf)} \log L(\theta) = \nabla^{(inf)} \log p(x|\theta) \tag{12c}$$

where $\nabla^{(inf)}$ denotes the gradient with respect to the information metric, and $L(\theta)$ is the likelihood ratio, equal to $p(x|\theta)$ divided by $p(x|\theta^*)$. Now, replacing (12) into (1a) defines an original algorithm for recursive estimation of the true scatter matrix θ^*.

To apply this algorithm in practice, one may evaluate $u(\theta, x)$ via the following steps. Denote $g(\theta, x)$ the gradient of $\log p(x|\theta)$ with respect to the affine-invariant metric of [29], which corresponds to $I_1 = 1$ and $I_2 = 0$. By direct calculation from (10), this is given by

$$g(\theta, x) \;=\; -\frac{1}{2}\theta - h'\left(x^\dagger \theta^{-1} x \right) x x^\dagger \tag{13a}$$

Moreover, introduce the constants $J_1 = I_1$ and $J_2 = I_1 + m I_2$. Then, $u(\theta, x)$ can be evaluated,

$$u(\theta, x) \;=\; J_1^{-1} \left(g(\theta, x) \right)^\perp \;+\; J_2^{-1} \left(g(\theta, x) \right)^\parallel \tag{13b}$$

from the orthogonal decomposition of $g = g(\theta, x)$,

$$g^\parallel \;=\; \text{tr}\left(\theta^{-1} g \right) \frac{\theta}{m} \qquad g^\perp \;=\; g - g^\parallel \tag{13c}$$

Figures 1 and 2 below display numerical results from an application to Kotz-type distributions, which correspond to $h(t) = -\frac{t^s}{2}$ in (10) and $\varphi = s^2 \frac{m}{2s} \left(\frac{m}{2s} + 1 \right)$ in (11c) [24,27]. These figures were generated from 10^3 Monte Carlo runs of the algorithm defined by (1a) and (12), with random initialisation, for the specific values $s = 4$ and $m = 7$. Essentially the same numerical results could be observed for any $s \leq 9$ and $m \leq 50$.

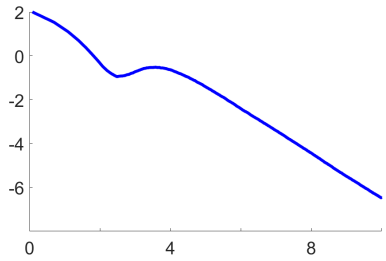

Figure 1. Fast non-asymptotic rate of convergence

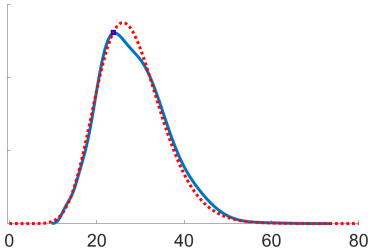

Figure 2. Asymptotic efficiency (optimal rate of convergence)

Figure 1 confirms the fast non-asymptotic rate of convergence (5), stated in (i) of Proposition 5. On a log-log scale, it shows the empirical mean $\mathbb{E}_{\mathrm{MC}}\, d^2(\theta_n, \theta^*)$ over Monte Carlo runs, as a function of n. This decreases with a constant negative slope equal to -1, starting roughly at $\log n = 4$. Here, the Riemannian distance $d(\theta_n, \theta^*)$ induced by the information metric (11) is given by [28]

$$d^2(\theta, \theta^*) \;=\; I_1 \,\mathrm{tr}\left[\log\left(\theta^{-1}\theta^*\right)\right]^2 \;+\; I_2 \,\mathrm{tr}^2\left[\log\left(\theta^{-1}\theta^*\right)\right] \qquad \theta\,, \theta^* \in \Theta \tag{14}$$

where log denotes the symmetric matrix logarithm [30]. Figure 2 confirms the asymptotic efficiency of the recursive estimates θ_n, stated in (iii) of Proposition 5, using Remark 4. It shows a kernel density estimate of $n\, d^2(\theta_n, \theta^*)$ where $n = 10^5$ (solid blue curve). This agrees with a χ^2-distribution with 28 degrees of freedom (dotted red curve), where $d = 28$ is indeed the dimension of the parameter space $\mathcal{P}_m$ for $m = 7$.

5. Conclusions

Recursive estimation is a subject that is over fifty years old [10], with its foundation in the general theory of stochastic optimisation [9,15]. Its applications are very wide-ranging, as they cover areas from control theory to machine learning [17].

With the increasing role of Riemannian manifolds in statistical inference and machine learning, it was natural to generalise the techniques of stochastic optimisation, from Euclidean space to Riemannian manifolds. Indeed, this started with the work of Bonnabel [1], which impulsed a series of successive works, such as [2,3].

These works have mostly sought to directly adapt classical results, known in Euclidean space, which concern optimal rates of convergence to a unique attractive minimum of a cost function. The present work also belongs to this line of thinking. It shows that when dealing with a recursive estimation problem, where the unknown statistical parameter belongs to a differentiable manifold, equipping this manifold with the information metric of the underlying statistical model, leads to optimal algorithm performance, which is moreover automatic (does not involve parameter tuning).

The results obtained in the present work (as well as in [2,3]) suffer from inherent limitations. Indeed, being only focused on convergence to a unique attractive minimum, it does not tackle the following important open problems :

- stability of stochastic optimisation algorithms finding verifiable conditions which ensure a stochastic optimisation algorithm remains within a compact set. A more general form of this problem is computing the probability of a stochastic optimisation algorithm exiting a certain neighborhood of a stationary point (whether attractive or not) within a finite number of iterations.
- non-asymptotic performance of stochastic optimisation algorithms: this involves computing explicitly the outcome which the algorithm is able to achieve, after a given finite number of iterations. This provides a much stronger theoretical guarantee, to the user, than standard results which compute a rate of convergence.

These problems have attracted much attention and generated well-known results when considered in the Euclidean case [31,32], but remain open in the context of Riemannian manifolds. They involve much richer interaction between Riemannian geometry and stochastic optimisation, due to their global nature.

6. Proofs of Main Results

Proof of Proposition 1. The proof is a generalisation of the original proof in [1], itself modeled on the proof for the Euclidean case in [33]. Throughout the following, let $\mathcal{X}_n$ be the σ-field generated by $x_1, \ldots, x_n$ [20]. Recall that $(x_n; n = 1, 2, \ldots)$ are i.i.d. with distribution P_{θ^*}. Therefore, by (1a), θ_n is $\mathcal{X}_n$-measurable and x_{n+1} is independent from $\mathcal{X}_n$. Thus, using elementary properties of conditional expectation [20],

$$\mathbb{E}\left[u(\theta_n, x_{n+1})|\mathcal{X}_n\right] = -D(\theta_n) \tag{15a}$$

$$\mathbb{E}\left[\|u(\theta_n, x_{n+1})\|^2 \middle| \mathcal{X}_n\right] = V(\theta_n) \tag{15b}$$

where (15a) follows from (1c), and (15b) from (u1). Let L be a Lipschitz constant for $\nabla D(\theta)$, and C be an upper bound on $V(\theta)$, for $\theta \in \Theta^*$. The following inequality is now proved, for any positive integer n,

$$\mathbb{E}\left[D(\theta_{n+1}) - D(\theta_n)|\mathcal{X}_n\right] \leq \gamma_{n+1}^2 LC - \gamma_{n+1}\|\nabla D(\theta_n)\|^2 \tag{16}$$

once this is done, Proposition 1 is obtained by applying the Robbins-Siegmund theorem [9].

Proof of (16): let $c(t)$ be the geodesic connecting θ_n to θ_{n+1} with equation

$$c(t) = \mathrm{Exp}_{\theta_n}(t\gamma_{n+1}u(\theta_n, x_{n+1})) \tag{17a}$$

From the fundamental theorem of calculus,

$$D(\theta_{n+1}) - D(\theta_n) = \gamma_{n+1}\left\langle u(\theta_n, x_{n+1}), \nabla D(\theta_n)\right\rangle + \gamma_{n+1}\int_0^1 \left[\langle \dot{c}, \nabla D\rangle_{c(t)} - \langle \dot{c}, \nabla D\rangle_{c(0)}\right] dt \tag{17b}$$

Since the recursive estimates θ_n are stable, θ_n and θ_{n+1} both lie in Θ^*. Since Θ^* is convex, the whole geodesic $c(t)$ lies in Θ^*. Then, since $\nabla D(\theta)$ is Lipschitz on Θ^*, it follows from (17b),

$$D(\theta_{n+1}) - D(\theta_n) \leq \gamma_{n+1}\left\langle u(\theta_n, x_{n+1}), \nabla D(\theta_n)\right\rangle + \gamma_{n+1}^2 L\|u(\theta_n, x_{n+1})\|^2 \tag{17c}$$

Taking conditional expectations in this inequality, and using (15a) and (15b),

$$\mathbb{E}\left[D(\theta_{n+1}) - D(\theta_n)|\mathcal{X}_n\right] \leq -\gamma_{n+1}\|\nabla D(\theta_n)\|^2 + \gamma_{n+1}^2 LV(\theta_n) \tag{17d}$$

so (16) follows since (u1) guarantees $V(\theta_n) \leq C$. □ *Conclusion*: by the Robbins-Siegmund theorem, inequality (16) implies that, almost surely,

$$\lim D(\theta_n) = D_\infty < \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} \gamma_{n+1}\|\nabla D(\theta_n)\|^2 < \infty \tag{18a}$$

In particular, from the first condition in (1b), convergence of the sum in (18a) implies

$$\lim \|\nabla D(\theta_n)\| = 0 \qquad \text{almost surely} \tag{18b}$$

Now, since the sequence of recursive estimates θ_n lies in the compact set Θ^*, it has at least one point of accumulation in this set, say θ_*. If $\theta_{n(k)}$ is a subsequence of θ_n, converging to θ_*,

$$\|\nabla D(\theta_*)\| = \lim \|\nabla D(\theta_{n(k)})\| = \lim \|\nabla D(\theta_n)\| = 0 \qquad \text{almost surely}$$

where the third equality follows from (18b). This means that θ_* is a stationary point of $D(\theta)$ in Θ^*. Thus, (d1) implies $\theta_* = \theta^*$ is the unique point of accumulation of θ_n. In other words, $\lim \theta_n = \theta^*$ almost surely. □

Proof of Proposition 2. The proof is modeled on the proofs for the Euclidean case, given in [10,15]. It relies on the following geometric Lemmas 1 and 2. Lemma 1 will be proved in Appendix A. On the other hand, Lemma 2 is the same as the trigonometric distance bound of [2]. For Lemma 1, recall that $\lambda > 0$ denotes the smallest eigenvalue of the matrix H defined in (4b).

Lemma 1. *for any $\mu < \lambda$, there exists a neighborhood $\bar{\Theta}^*$ of θ^*, contained in Θ^*, with*

$$\langle \mathrm{Exp}_\theta^{-1}(\theta^*), \nabla D(\theta) \rangle \leq -\mu\, d^2(\theta, \theta^*) \qquad \text{for } \theta \in \bar{\Theta}^* \tag{19a}$$

Lemma 2. *let $-\kappa^2$ be a lower bound on the sectional curvature of Θ in Θ^*, and $C_\kappa = R\kappa \coth(R\kappa)$ where R is the diameter of Θ^*. For $\tau, \theta \in \Theta^*$, where $\tau = \mathrm{Exp}_\theta(u)$,*

$$d^2(\tau, \theta^*) \leq d^2(\theta, \theta^*) - 2\,\langle \mathrm{Exp}_\theta^{-1}(\theta^*), u \rangle + C_\kappa \|u\|^2 \tag{19b}$$

Proof of (5): let $\gamma_n = \frac{a}{n}$ with $2\lambda a > 2\mu a > 1$ for some $\mu < \lambda$, and let $\bar{\Theta}^*$ be the neighborhood corresponding to μ in Lemma 1. By Proposition 1, the θ_n converge to θ^* almost surely. Without loss of generality, it can be assumed that all the θ_n lie in $\bar{\Theta}^*$, almost surely. Then, (1a) and Lemma 2 imply, for any positive integer n,

$$d^2(\theta_{n+1}, \theta^*) \leq d^2(\theta_n, \theta^*) - 2\gamma_{n+1}\,\langle \mathrm{Exp}_{\theta_n}^{-1}(\theta^*), u(\theta_n, x_{n+1}) \rangle + \gamma_{n+1}^2\, C_\kappa \|u(\theta_n, x_{n+1})\|^2 \tag{20a}$$

Indeed, this follows by replacing $\tau = \theta_{n+1}$ and $\theta = \theta_n$ in (19b). Taking conditional expectations in (20a), and using (15a) and (15b),

$$\mathbb{E}\,[d^2(\theta_{n+1}, \theta^*) | \mathcal{X}_n] \leq d^2(\theta_n, \theta^*) + 2\gamma_{n+1}\,\langle \mathrm{Exp}_{\theta_n}^{-1}(\theta^*), \nabla D(\theta_n) \rangle + \gamma_{n+1}^2\, C_\kappa V(\theta_n)$$

Then, by (u1) and (19a) of Lemma 1,

$$\mathbb{E}\,[d^2(\theta_{n+1}, \theta^*) | \mathcal{X}_n] \leq d^2(\theta_n, \theta^*)(1 - 2\gamma_{n+1}\mu) + \gamma_{n+1}^2\, C_\kappa C \tag{20b}$$

where C is an upper bound on $V(\theta)$, for $\theta \in \Theta^*$. By further taking expectations

$$\mathbb{E}\, d^2(\theta_{n+1}, \theta^*) \leq \mathbb{E}\, d^2(\theta_n, \theta^*)(1 - 2\gamma_{n+1}\mu) + \gamma_{n+1}^2\, C_\kappa C \tag{20c}$$

Using (20c), the proof reduces to an elementary reasoning by recurrence. Indeed, replacing $\gamma_n = \frac{a}{n}$ into (20c), it follows that

$$\mathbb{E}\, d^2(\theta_{n+1}, \theta^*) \leq \mathbb{E}\, d^2(\theta_n, \theta^*) \left(1 - \frac{2\mu a}{n+1}\right) + \frac{a^2 C_\kappa C}{(n+1)^2} \tag{21a}$$

On the other hand, if $b(n) = \frac{b}{n}$ where $b > a^2 C_\kappa C \, (2\mu a - 1)^{-1}$, then

$$b(n+1) \geq b(n) \left(1 - \frac{2\mu a}{n+1}\right) + \frac{a^2 C_\kappa C}{(n+1)^2} \tag{21b}$$

Let b be sufficiently large, so (21b) is verified and $\mathbb{E}\, d^2(\theta_{n_o}, \theta^*) \leq b(n_o)$ for some n_o. Then, by recurrence, using (21a) and (21b), one also has that $\mathbb{E}\, d^2(\theta_n, \theta^*) \leq b(n)$ for all $n \geq n_o$. In other words, (5) holds true. □

Proof of Proposition 3. the proof is modeled on the proof for the Euclidean case in [10]. To begin, let W_n be the stochastic process given by

$$W_n \;=\; n^p \, d^2(\theta_n, \theta^*) + n^{-q} \qquad \text{where } q \in (0, 1-p) \tag{22a}$$

The idea is to show that this process is a positive supermartingale, for sufficiently large n. By the supermartingale convergence theorem [20], it then follows that W_n converges to a finite limit, almost surely. In particular, this implies

$$\lim n^p \, d^2(\theta_n, \theta^*) \;=\; \ell_p < \infty \qquad \text{almost surely} \tag{22b}$$

Then, ℓ_p must be equal to zero, since p is arbitrary in the interval $(0,1)$. Precisely, for any $\varepsilon \in (0, 1-p)$,

$$\ell_p \;=\; \lim n^p \, d^2(\theta_n, \theta^*) \;=\; \lim n^{-\varepsilon} n^{p+\varepsilon} \, d^2(\theta_n, \theta^*) \;=\; (\lim n^{-\varepsilon}) \; \ell_{p+\varepsilon} \;=\; 0$$

It remains to show that W_n is a supermartingale, for sufficiently large n. To do so, note that by (20b) from the proof of Proposition 2,

$$\mathbb{E}\,[W_{n+1} - W_n | \mathcal{X}_n] \leq \; d^2(\theta_n, \theta^*) \, \frac{p - 2\mu a}{(n+1)^{1-p}} \; + \; \frac{a^2 C_\kappa C}{(n+1)^{2-p}} \; - \; \frac{q}{(n+1)^{q+1}}$$

Here, the first term on the right-hand side is negative, since $2\mu a > 1 > p$. Moreover, the third term dominates the second one for sufficiently large n, since $q < 1 - p$. Thus, for sufficiently large n, the right-hand side is negative, and W_n is a supermartingale. □

Proof of Proposition 4. the proof relies on the following geometric Lemmas 3 and 4, which are used to linearise Algorithm (1a), in terms of the normal coordinates θ^α. This idea of linearisation in terms of local coordinates also plays a central role in [3].

Lemma 3. *let θ_n, θ_{n+1} be given by (1a) with $\gamma_n = \frac{a}{n}$. Then, in a system of normal coordinates with origin at θ^*,*

$$\theta^\alpha_{n+1} \;=\; \theta^\alpha_n + \gamma_{n+1} \, u^\alpha_{n+1} + \gamma^2_{n+1} \, \pi^\alpha_{n+1} \qquad \mathbb{E}\,|\pi^\alpha_{n+1}| = O(n^{-1/2}) \tag{23a}$$

where u^α_{n+1} are the components of $u(\theta_n, x_{n+1})$.

Lemma 4. *let $v_n = \nabla D(\theta_n)$. Then, in a system of normal coordinates with origin at θ^*,*

$$v^\alpha_n \;=\; H_{\alpha\beta} \, \theta^\beta_n \; + \; \rho^\alpha_n \qquad \rho^\alpha_n = o\left(d(\theta_n, \theta^*)\right) \tag{23b}$$

where v^α_n are the components of v_n and the $H_{\alpha\beta}$ were defined in (4b).

Linearisation of (1a): let $u(\theta_n, x_{n+1}) = -v_n + w_{n+1}$. Then, it follows from (23a) and (23b),

$$\theta^\alpha_{n+1} \;=\; \theta^\alpha_n \; - \; \gamma_{n+1} \, H_{\alpha\beta} \, \theta^\beta_n \; - \; \gamma_{n+1} \, \rho^\alpha_n \; + \; \gamma_{n+1} \, w^\alpha_{n+1} \; + \; \gamma^2_{n+1} \, \pi^\alpha_{n+1} \tag{24a}$$

Denote the re-scaled coordinates $n^{1/2}\theta_n^\alpha$ by η_n^α, and recall $\gamma_n = \frac{a}{n}$. Then, using the estimate $(n+1)^{1/2} = n^{1/2}(1 + (2n)^{-1} + O(n^{-2}))$, it follows from (24a) that

$$\eta_{n+1}^\alpha \;=\; \eta_n^\alpha + \frac{A_{\alpha\beta}}{n+1}\,\eta_n^\beta \;+\; \frac{a}{(n+1)^{1/2}}\left[B_{\alpha\beta}\,\theta_n^\beta \;-\rho_n^\alpha \;+\; w_{n+1}^\alpha \;+\; \frac{a\pi_{n+1}^\alpha}{n+1}\right] \tag{24b}$$

where $A_{\alpha\beta} = \frac{1}{2}\delta_{\alpha\beta} - aH_{\alpha\beta}$ and $B_{\alpha\beta} = O(n^{-1})$. Equation (24b) is a first-order, inhomogeneous, linear difference equation, for the "vector" η_n of components η_n^α. □

Study of equation (24b) : switching to vector-matrix notation, equation (24b) is of the general form

$$\eta_{n+1} \;=\; \left(I \;+\; \frac{A}{n+1}\right)\,\eta_n \;+\; \frac{a\,\xi_{n+1}}{(n+1)^{1/2}} \tag{25a}$$

where I denotes the identity matrix, A has matrix elements $A_{\alpha\beta}$, and (ξ_n) is a sequence of inputs. The general solution of this equation is [10,34]

$$\eta_n \;=\; A_{n,m}\,\eta_m \;+\; \sum_{k=m+1}^{n} A_{n,k}\,\frac{a\,\xi_k}{k^{1/2}} \qquad \text{for} \;\; n \geq m \tag{25b}$$

where the transition matrix $A_{n,k}$ is given by

$$A_{n,k} \;=\; \prod_{j=k+1}^{n}\left(I + \frac{A}{j}\right) \qquad A_{n,n} = I \tag{25c}$$

Since $2\lambda a > 1$, the matrix A is stable. This can be used to show that [10,34]

$$q > \frac{1}{2} \;\text{ and }\; \mathbb{E}\,|\xi_n| = O(n^{-q}) \;\Longrightarrow\; \lim \eta_n \;=\; 0 \;\text{ in probability} \tag{25d}$$

where $|\xi_n|$ denotes the Euclidean vector norm. Then, it follows from (25d) that η_n converges to zero in probability, in each one of the three cases

$$\xi_{n+1}^\alpha \;=\; B_{\alpha\beta}\,\theta_n^\beta \;\;;\;\; \xi_{n+1}^\alpha \;=\; \rho_n^\alpha \;\;;\;\; \xi_{n+1}^\alpha \;=\; \frac{\pi_{n+1}^\alpha}{n+1}$$

Indeed, in the first two cases, the condition required in (25d) can be verified using (5), whereas in the third case, it follows immediately from the estimate of $\mathbb{E}|\pi_{n+1}^\alpha|$ in (23a). □

Conclusion : by linearity of (24b), it is enough to consider the case $\xi_{n+1}^\alpha = w_{n+1}^\alpha$ in (25a). Then, according to (25b), η_n has the same limit distribution as the sums

$$\tilde{\eta}_n \;=\; \sum_{k=1}^{n} A_{n,k}\,\frac{a w_k}{k^{1/2}} \tag{26}$$

By (15), (w_k) is a sequence of square-integrable martingale differences. Therefore, to conclude that the limit distribution of $\tilde{\eta}_n$ is a centred d-variate normal distribution, with covariance matrix Σ given by (7), it is enough to verify the conditions of the martingale central limit theorem [35],

$$\lim \max_{k\leq n}\left|A_{n,k}\,\frac{a w_k}{k^{1/2}}\right| \;=\; 0 \;\text{ in probability} \tag{27a}$$

$$\sup\,\mathbb{E}\,|\tilde{\eta}_n|^2 \;<\; \infty \tag{27b}$$

$$\lim \sum_{k=1}^{n} \frac{a^2}{k}\,A_{n,k}\,\Sigma_k\,A_{n,k} \;=\; \Sigma \;\text{ in probability} \tag{27c}$$

where Σ_k is the conditional covariance matrix

$$\Sigma_k \;=\; \mathbb{E}\left[w_k w_k^\dagger \middle| \mathcal{X}_{k-1}\right] \tag{28}$$

Conditions (27) are verified in Appendix B, which completes the proof. □

Proof of Proposition 5. Denote $\partial_\alpha = \frac{\partial}{\partial \theta^\alpha}$ the coordinate vector fields of the normal coordinates θ^α. Since $\langle \cdot, \cdot \rangle$ coincides with the information metric of the model P, it follows from (4b) and (A10),

$$H_{\alpha\beta} \;=\; \langle \partial_\alpha , \partial_\beta \rangle_{\theta^*} \tag{29a}$$

However, by the definition of normal coordinates [16], the ∂_α are orthonormal at θ^*. Therefore,

$$H_{\alpha\beta} \;=\; \delta_{\alpha\beta} \tag{29b}$$

Thus, the matrix H is equal to the identity matrix, and its smallest eigenvalue is $\lambda = 1$.

Proof of (i): this follows directly from Propositions 2 and 3. Indeed, since $\lambda = 1$, the conditions of these propositions are verified, as soon as $2a > 1$. Therefore, (5) and (6) hold true. □

Proof of (ii): this follows from Proposition 4. The conditions of this proposition are verified, as soon as $2a > 1$. Therefore, the distribution of the re-scaled coordinates $(n^{1/2}\theta_n^\alpha)$ converges to a centred d-variate normal distribution, with covariance matrix Σ given by Lyapunov's equation (7). If $a = 1$, then (29b) implies $A_{\alpha\beta} = -\frac{1}{2}\delta_{\alpha\beta}$, so that Lyapunov's equation (7) reads $\Sigma = \Sigma^*$, as required. □

For the following proof of (iii), the reader may wish to recall that summation convention is used throughout the present work. That is [16], summation is implicitly understood over any repeated subscript or superscript from the Greek alphabet, taking the values $1, \ldots, d$.

Proof of (iii): let $\ell(\theta) = \log L(\theta)$ and assume $u(\theta, x)$ is given by (2b). Then, by the definition of normal coordinates [16], the following expression holds

$$u^\alpha(\theta^*) \;=\; \left.\frac{\partial \ell}{\partial \theta^\alpha}\right|_{\theta^\alpha = 0} \tag{30a}$$

Replacing this into (4a) gives

$$\Sigma^*_{\alpha\beta} \;=\; E_{\theta^*}\left[\frac{\partial \ell}{\partial \theta^\alpha}\frac{\partial \ell}{\partial \theta^\beta}\right]_{\theta^\alpha = 0} \;=\; -\,E_{\theta^*}\left.\frac{\partial^2 \ell}{\partial \theta^\alpha \partial \theta^\beta}\right|_{\theta^\alpha = 0} \;=\; \left.\frac{\partial^2 D}{\partial \theta^\alpha \partial \theta^\beta}\right|_{\theta^\alpha = 0} \tag{30b}$$

where the second equality is the so-called Fisher's identity (see [19], Page 28), and the third equality follows from (2a) by differentiating under the expectation. Now, by (4b) and (29b), Σ^* is the identity matrix.

To show that the recursive estimates θ_n are asymptotically efficient, let $(\tau^\alpha ; \alpha = 1, \ldots, d)$ be any local coordinates with origin at θ^* and let $\tau_n^\alpha = \tau^\alpha(\theta_n)$. From the second-order Taylor expansion of each coordinate function τ^α, it is straightforward to show that

$$n^{1/2}\tau_n^\alpha \;=\; \left(\frac{\partial \tau^\alpha}{\partial \theta^\gamma}\right)_{\theta^*}\left(n^{1/2}\theta_n^\gamma\right) \;+\; \sigma^\alpha(\theta_n)\left(n^{1/2} d^2(\theta_n, \theta^*)\right) \tag{31a}$$

where the subscript θ^* indicates the derivative is evaluated at θ^*, and where σ^α is a continuous function in the neighborhood of θ^*. By (6), the second term in (31a) converges to zero almost surely. Therefore, the limit distribution of the re-scaled coordinates $(n^{1/2}\tau_n^\alpha)$ is the same as that of the first term in (31a). By (ii), this is a centred d-variate normal distribution with covariance matrix Σ^τ given by

$$\Sigma^\tau_{\alpha\beta} \;=\; \left(\frac{\partial \tau^\alpha}{\partial \theta^\gamma}\right)_{\theta^*} \Sigma^*_{\gamma\kappa} \left(\frac{\partial \tau^\beta}{\partial \theta^\kappa}\right)_{\theta^*} \;=\; \left(\frac{\partial \tau^\alpha}{\partial \theta^\gamma}\right)_{\theta^*}\left(\frac{\partial \tau^\beta}{\partial \theta^\gamma}\right)_{\theta^*} \tag{31b}$$

where the second equality follows because $\Sigma^*_{\gamma\kappa} = \delta_{\gamma\kappa}$ since Σ^* is the identity matrix.

It remains to show that Σ^τ is the inverse of the information matrix I^τ as in (A12). According to (A10), this is given by

$$I^\tau_{\alpha\beta} \;=\; \frac{\partial^2 D}{\partial\tau^\alpha \partial\tau^\beta}\bigg|_{\tau^\alpha=0} \;=\; -\,E_{\theta^*}\,\frac{\partial^2 \ell}{\partial\tau^\alpha \partial\tau^\beta}\bigg|_{\tau^\alpha=0} \;=\; E_{\theta^*}\left[\frac{\partial \ell}{\partial\tau^\alpha}\frac{\partial \ell}{\partial\tau^\beta}\right]_{\tau^\alpha=0} \tag{31c}$$

where the second equality follows from (2a), and the third equality from Fisher's identity (see [19], Page 28). Now, a direct application of the chain rule yields the following

$$I^\tau_{\alpha\beta} \;=\; E_{\theta^*}\left[\frac{\partial \ell}{\partial\tau^\alpha}\frac{\partial \ell}{\partial\tau^\beta}\right]_{\tau^\alpha=0} \;=\; \left(\frac{\partial\theta^\gamma}{\partial\tau^\alpha}\right)_{\theta^*} E_{\theta^*}\left[\frac{\partial \ell}{\partial\theta^\gamma}\frac{\partial \ell}{\partial\theta^\kappa}\right]_{\theta^\gamma=0}\left(\frac{\partial\theta^\kappa}{\partial\tau^\beta}\right)_{\theta^*}$$

By the first equality in (30b), this is equal to

$$I^\tau_{\alpha\beta} \;=\; \left(\frac{\partial\theta^\gamma}{\partial\tau^\alpha}\right)_{\theta^*} \Sigma^*_{\gamma\kappa}\left(\frac{\partial\theta^\kappa}{\partial\tau^\beta}\right)_{\theta^*} \;=\; \left(\frac{\partial\theta^\gamma}{\partial\tau^\alpha}\right)_{\theta^*}\left(\frac{\partial\theta^\gamma}{\partial\tau^\beta}\right)_{\theta^*} \tag{31d}$$

because $\Sigma^*_{\gamma\kappa} = \delta_{\gamma\kappa}$ is the identity matrix. Comparing (31b) to (31d), it is clear that Σ^τ is the inverse of the information matrix I^τ as in (A12).

Proof of (iv): (8a) and (8b) follow from (5) and (6), respectively, by using (A11). Precisely, it is possible to write (A11) in the form

$$D(\theta_n) \;=\; \frac{1}{2}\,d^2(\theta_n,\theta^*) \;+\; \omega(\theta_n)\,d^2(\theta_n,\theta^*) \tag{32a}$$

where ω is a continuous function in the neighborhood of θ^*, equal to zero at $\theta = \theta^*$. To obtain (8a), it is enough to take expectations in (32a) and note that ω is bounded above in the neighborhood of θ^*. Then, (8a) follows directly from (5).

To obtain (8b), it is enough to multiply (32a) by n^p where $p \in (0,1)$. This gives the following expression

$$n^p D(\theta_n) \;=\; \frac{1}{2}\,n^p d^2(\theta_n,\theta^*)\,(1+\omega(\theta_n)) \tag{32b}$$

From (6), $n^p d^2(\theta_n,\theta^*)$ converges to zero almost surely. Moreover, by continuity of ω, it follows that $\omega(\theta_n)$ converges to $\omega(\theta^*) = 0$ almost surely. Therefore, by taking limits in (32b), it is readily seen that

$$\lim\, n^p D(\theta_n) \;=\; \frac{1}{2}\,(\lim\, n^p d^2(\theta_n,\theta^*))\,(1+\lim\,\omega(\theta_n)) \;=\; 0 \tag{32c}$$

almost surely. However, this is equivalent to the statement that $D(\theta_n) = o(n^{-p})$ for $p \in (0,1)$, almost surely. Thus, (8b) is proved. □

Author Contributions: Data curation, J.Z.; Software, J.Z.; Writing-original draft, S.S.; Writing-review & editing, S.S.

Funding: Agence Nationale de la Recherche : ANR-17-ASTR-0015

Acknowledgments: At the end, we thank the support from the ANR MARGARITA (Modern Adaptive Radar: Great Advances in Robust and Inference Techniques and Application) under Grant ANR-17-ASTR-0015 for our works.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs of Geometric Lemmas

Appendix A.1. Lemma 1

Let $c(t)$ be the geodesic connecting θ^* to some $\theta \in \Theta^*$, parameterised by arc length. In other words, $c(0) = \theta^*$ and $c(t_\theta) = \theta$ where $t_\theta = d(\theta, \theta^*)$. Denote Π_t the parallel transport along $c(t)$, from $T_{c(0)}\Theta$ to $T_{c(t)}\Theta$. Since the velocity $\dot{c}(t)$ is self-parallel [16],

$$\dot{c}(t_\theta) \;=\; \Pi_{t_\theta}(\dot{c}(0))$$

Multiplying this identity by $-t_\theta$, it follows that

$$\mathrm{Exp}_\theta^{-1}(\theta^*) \;=\; -\,\Pi_{t_\theta}\big(\mathrm{Exp}_{\theta^*}^{-1}(\theta)\big) \tag{A1a}$$

Moreover, recall the first-order Taylor expansion of the gradient $\nabla D(\theta)$ [16,36]

$$\nabla D(\theta) \;=\; \Pi_{t_\theta}\Big(\nabla D(\theta^*) + t_\theta\,\nabla^2 D(\theta^*)\cdot\dot{c}(0) + t_\theta\,\phi(\theta)\Big) \tag{A1b}$$

where $\phi(\theta)$ is continuous and equal to zero at $\theta = \theta^*$. Here, $\nabla^2 D(\theta^*)$ is the Hessian of $D(\theta)$ at $\theta = \theta^*$, considered as a linear mapping of $T_{\theta^*}\Theta$ [16,36]

$$\nabla^2 D(\theta^*)\cdot w \;=\; \nabla_w \nabla D(\theta^*) \qquad \text{for } \; w \in T_{\theta^*}\Theta$$

where ∇_w denotes the covariant derivative in the direction of w. By (d1), the first term on the right-hand side of (A1b) is equal to zero, so that

$$\nabla D(\theta) \;=\; \Pi_{t_\theta}\Big(\nabla^2 D(\theta^*)\cdot\mathrm{Exp}_{\theta^*}^{-1}(\theta) + t_\theta\,\phi(\theta)\Big) \tag{A1c}$$

Taking the scalar product of (A1a) and (A1c),

$$\langle \mathrm{Exp}_\theta^{-1}(\theta^*), \nabla D(\theta)\rangle \;=\; -\,\langle \mathrm{Exp}_{\theta^*}^{-1}(\theta), \nabla^2 D(\theta^*)\cdot\mathrm{Exp}_{\theta^*}^{-1}(\theta)\rangle \;-\; t_\theta\,\langle \mathrm{Exp}_{\theta^*}^{-1}(\theta), \phi(\theta)\rangle \tag{A1d}$$

since parallel transport preserves scalar products. In terms of the normal coordinates θ^α, this reads [16]

$$\langle \mathrm{Exp}_\theta^{-1}(\theta^*), \nabla D(\theta)\rangle \;=\; -\,H_{\alpha\beta}\,\theta^\alpha\theta^\beta \;-\; t_\theta^2\;\hat{\theta}^\alpha\phi^\alpha \tag{A1e}$$

where $H = (H_{\alpha\beta})$ was defined in (4b), $\hat{\theta}^\alpha$ denotes the quotient θ^α / t_θ, and the ϕ^α denote the components of $\phi(\theta)$. Note that $t_\theta^2 = d^2(\theta, \theta^*) = \theta^\alpha\theta^\alpha$, so (A1e) can be written

$$\langle \mathrm{Exp}_\theta^{-1}(\theta^*), \nabla D(\theta)\rangle \;=\; (\psi(\theta)\delta_{\alpha\beta} \;-\; H_{\alpha\beta})\;\theta^\alpha\theta^\beta \tag{A1f}$$

where $\psi(\theta)$ is continuous and equal to zero at $\theta = \theta^*$. To conclude, let $\mu = \lambda - \varepsilon$ for some $\varepsilon > 0$, and $\bar{\Theta}^*$ a neighborhood of θ^*, contained in Θ^*, such that $\psi(\theta) \le \varepsilon$ for $\theta \in \bar{\Theta}^*$. Then, since λ is the smallest eigenvalue of $H = (H_{\alpha\beta})$,

$$\langle \mathrm{Exp}_\theta^{-1}(\theta^*), \nabla D(\theta)\rangle \;\le\; (\varepsilon - \lambda)\;\theta^\alpha\theta^\alpha \;=\; -\mu\, d^2(\theta, \theta^*)$$

for $\theta \in \bar{\Theta}^*$. This is exactly (19a), so the lemma is proved. □

Appendix A.2. Lemma 3

To simplify notation, let $u_{n+1} = u(\theta_n, x_{n+1})$. Then, the geodesic $c(t)$, connecting θ_n to θ_{n+1}, has equation

$$c(t) \;=\; \mathrm{Exp}_{\theta_n}(t\gamma_{n+1}u_{n+1})$$

Each one of the normal coordinates θ^α is a C^3 function $\theta^\alpha : \Theta^* \to \mathbb{R}$, with differential $d\theta^\alpha$ and Hessian [16]

$$\nabla^2 \theta^\alpha \;=\; -\Gamma^\alpha_{\beta\gamma}(\theta)\, d\theta^\beta \otimes d\theta^\gamma$$

where $\Gamma^\alpha_{\beta\gamma}$ are the Christoffel symbols of the coordinates θ^α, and $\otimes$ denotes the tensor product. Then, the second-order Taylor expansion of the functions $\theta^\alpha \circ c$ reads

$$(\theta^\alpha \circ c)(1) = (\theta^\alpha \circ c)(0) + \gamma_{n+1}\, d\theta^\alpha(u_{n+1}) - \frac{1}{2}\gamma_{n+1}^2\, \Gamma^\alpha_{\beta\gamma}(\theta_n)\, d\theta^\beta(u_{n+1})\, d\theta^\gamma(u_{n+1}) + \gamma_{n+1}^3 T^\alpha_{n+1} \tag{A2a}$$

where T^α_{n+1} satisfies

$$|T^\alpha_{n+1}| \;\leq\; K_1 \,\|u_{n+1}\|^3 \tag{A2b}$$

for a constant K_1 which does not depend on n, as can be shown by direct calculation. Of course, $(\theta^\alpha \circ c)(1) = \theta^\alpha_{n+1}$ and $(\theta^\alpha \circ c)(0) = \theta^\alpha_n$. Moreover, $d\theta^\alpha(u_{n+1}) = u^\alpha_{n+1}$ are the components of u_{n+1}. Replacing into (A2a), this yields

$$\theta^\alpha_{n+1} \;=\; \theta^\alpha_n \;+\; \gamma_{n+1}\, u^\alpha_{n+1} \;+\; \gamma^2_{n+1}\, \pi^\alpha_{n+1} \tag{A2c}$$

where π^α_{n+1} is given by

$$\pi^\alpha_{n+1} \;=\; \gamma_{n+1}\, T^\alpha_{n+1} \;-\; \frac{1}{2}\, \Gamma^\alpha_{\beta\gamma}(\theta_n)\, u^\beta_{n+1}\, u^\gamma_{n+1} \tag{A2d}$$

Comparing (A2c) to (23a), it is clear the proof will be complete upon showing $\mathbb{E}\,|\pi^\alpha_{n+1}| = O(n^{-1/2})$. To do so, note that each Christoffel symbol $\Gamma^\alpha_{\beta\gamma}$ is a C^1 function on the compact set Θ^*, with $\Gamma^\alpha_{\beta\gamma}(\theta^*) = 0$ by the definition of normal coordinates [16]. Therefore,

$$\left|\, \Gamma^\alpha_{\beta\gamma}(\theta) \,\right| \;\leq\; K_2\, d(\theta, \theta^*) \tag{A2e}$$

for a constant K_2 which does not depend on n. Replacing the inequalities (A2b) and (A2e) into (A2d), and taking expectations, it follows that

$$\mathbb{E}\,|\pi^\alpha_{n+1}| \;\leq\; \gamma_{n+1}\, K_1\, \mathbb{E}\,\|u_{n+1}\|^3 \;+\; d^2 \times K_2\, \mathbb{E}\left[\, d(\theta_n, \theta^*)\, \|u_{n+1}\|^2 \,\right] \tag{A3a}$$

where d is the dimension of the parameter space Θ. However, using the fact that the x_n are i.i.d. with distribution P_{θ^*},

$$\mathbb{E}\left[\, \|u_{n+1}\|^3 \,\middle|\, \mathcal{X}_n \,\right] \;=\; E_{\theta^*}\, \|u(\theta_n, x)\|^3 \;\leq\; R^{3/4}(\theta_n) \tag{A3b}$$

by (u2) and Jensen's inequality [20]. On the other hand, by the Cauchy–Schwarz inequality,

$$\mathbb{E}\left[\, d(\theta_n, \theta^*)\, \|u_{n+1}\|^2 \,\right] \;\leq\; \left(\mathbb{E}\, d^2(\theta_n, \theta^*)\right)^{1/2} \left(\mathbb{E}\,\|u_{n+1}\|^4\right)^{1/2} \;\leq\; b\, n^{-1/2} \left(\mathbb{E}\,\|u_{n+1}\|^4\right)^{1/2}$$

for some $b > 0$ as follows from (5). Then, by the same reasoning that lead to (A3b),

$$\mathbb{E}\left[\, d(\theta_n, \theta^*)\, \|u_{n+1}\|^2 \,\right] \;\leq\; b\, n^{-1/2} \left(\mathbb{E}\, R(\theta_n)\right)^{1/2} \tag{A3c}$$

By (u2), there exists a uniform upper bound M on $R(\theta)$ for $\theta \in \Theta^*$. Since θ_n lies in Θ^* for all n, it follows by replacing the inequalities (A3b) and (A3c) into (A3a) that

$$\mathbb{E}\,|\pi^\alpha_{n+1}| \;\leq\; \gamma_{n+1}\, K_1\, M^{3/4} \;+\; d^2 \times K_2\, b\, n^{-1/2} M^{1/2} \tag{A3d}$$

Finally, by recalling that $\gamma_n = \frac{a}{n}$, it is clear that the right-hand side of (A3d) is $O(n^{-1/2})$, so the proof is complete. □

Appendix A.3. Lemma 4

The lemma is an instance of the general statement: let $\theta \in \Theta^*$ and $v = \nabla D(\theta)$. Then, in a system of normal coordinates with origin at θ^*,

$$v^\alpha \;=\; H_{\alpha\beta}\,\theta^\beta \;+\; o\left(d(\theta,\theta^*)\right) \tag{A4a}$$

where v^α are the components of v. Indeed, (23b) follows from (A4a) after replacing $\theta = \theta_n$, so that $v = v_n$, and setting

$$\rho_n^\alpha \;=\; v_n^\alpha \;-\; H_{\alpha\beta}\,\theta_n^\beta$$

To prove (A4a), recall (A1c) from the proof of Lemma 1, which can be written

$$v \;=\; \Pi_{t_\theta}\Big(\nabla^2 D(\theta^*)\cdot \mathrm{Exp}_{\theta^*}^{-1}(\theta)\Big) \;+\; d(\theta,\theta^*)\,\Pi_{t_\theta}(\phi(\theta)) \tag{A4b}$$

Denote $\partial_\alpha = \frac{\partial}{\partial\theta^\alpha}$ the coordinate vector fields of the normal coordinates θ^α. Note that [16,36]

$$\mathrm{Exp}_{\theta^*}^{-1}(\theta) \;=\; \theta^\beta\,\partial_\beta(\theta^*) \qquad\qquad \nabla^2 D(\theta^*)\cdot\partial_\beta(\theta^*) \;=\; H_{\alpha\beta}\,\partial_\alpha(\theta^*)$$

Replacing in (A4b), this gives

$$v \;=\; H_{\alpha\beta}\,\theta^\beta\;\Pi_{t_\theta}(\partial_\alpha(\theta^*)) \;+\; d(\theta,\theta^*)\,\Pi_{t_\theta}(\phi(\theta)) \tag{A4c}$$

From the first-order Taylor expansion of the vector fields ∂_α [16,36]

$$\partial_\alpha(\theta) \;=\; \Pi_{t_\theta}\big(\partial_\alpha(\theta^*) \;+\; \nabla\partial_\alpha(\theta^*)\cdot\mathrm{Exp}_{\theta^*}^{-1}(\theta)\big) \;+\; d(\theta,\theta^*)\,\Pi_{t_\theta}(\chi^\alpha(\theta))$$

where $\chi^\alpha(\theta)$ is continuous and equal to zero at $\theta = \theta^*$. However, by the definition of normal coordinates [16], each covariant derivative $\nabla\partial_\alpha(\theta^*)$ is zero. In other words,

$$\partial_\alpha(\theta) \;=\; \Pi_{t_\theta}(\partial_\alpha(\theta^*)) \;+\; d(\theta,\theta^*)\,\Pi_{t_\theta}(\chi^\alpha(\theta)) \tag{A4d}$$

Replacing (A4d) into (A4c), it follows

$$v \;=\; H_{\alpha\beta}\,\theta^\beta\,\partial_\alpha(\theta) \;+\; d(\theta,\theta^*)\,\Pi_{t_\theta}\big(\phi(\theta) - H_{\alpha\beta}\,\theta^\beta\chi^\alpha(\theta)\big) \tag{A4e}$$

Now, to obtain (A4a), it is enough to note the decomposition $v = v^\alpha\,\partial_\alpha(\theta)$ is unique, and $\phi(\theta) - H_{\alpha\beta}\,\theta^\beta\chi^\alpha(\theta)$ converges to zero as θ converges to θ^*. □

Appendix B. Conditions of the Martingale CLT

For the verification of Conditions (27), the following inequality (A5) will be useful. Let $\nu = a\lambda - \frac{1}{2}$, so $-\nu$ is the largest eigenvalue of the matrix A in (25a). There exists a constant C_A such that the transition matrices $A_{n,k}$ in (25c) satisfy [10,34]

$$|A_{n,k}|_{\mathrm{Op}} \;\le C_A\left(\frac{k}{n}\right)^\nu \tag{A5}$$

where $|A_{n,k}|_{\mathrm{Op}}$ denotes the Euclidean operator norm, equal to the largest singular value of the matrix $A_{n,k}$.

Condition (27a): to verify this condition, note that for arbitrary $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{k\le n}\left|A_{n,k}\,\frac{aw_k}{k^{1/2}}\right| > \varepsilon\right) \;\le \sum_{k=1}^n \mathbb{P}\left(\left|A_{n,k}\,\frac{aw_k}{k^{1/2}}\right| > \varepsilon\right) \;\le \sum_{k=1}^n \mathbb{P}\left(C_A\left(\frac{k}{n}\right)^\nu \left|\frac{aw_k}{k^{1/2}}\right| > \varepsilon\right) \tag{A6a}$$

where the second inequality follows from (A5). However, it follows from (u2) that there exists a uniform upper bound M_w on the fourth-order moments of $|w_k|$. Therefore, by Chebyshev's inequality [20]

$$\sum_{k=1}^{n} \mathbb{P}\left(C_A \left(\frac{k}{n}\right)^{\nu} \left| \frac{a w_k}{k^{1/2}} \right| > \varepsilon \right) \;\leq\; \left(\frac{a C_A}{\varepsilon}\right)^4 \frac{M_w}{n^{4\nu}} \sum_{k=1}^{n} k^{4\nu-2} \tag{A6b}$$

Since $\nu > 0$, the right-hand side of (A6b) has limit equal to 0 as $n \to \infty$, by the Euler–Maclaurin formula [37]. Replacing this limit from (A6b) into (A6a) immediately yields Condition (27a). □

Condition (27b): to verify this condition, recall that (w_k) is a sequence of square-integrable martingale differences. Therefore, from (26)

$$\mathbb{E}\,|\tilde{\eta}_n|^2 \;=\; \sum_{k=1}^{n} \frac{a^2}{k}\, \mathbb{E}\,\mathrm{tr}\big(A_{n,k}^2 \Sigma_k\big) \tag{A7a}$$

where Σ_k is the conditional covariance matrix in (28). Applying (A5) to each term under the sum in (A7a), it follows that

$$\mathbb{E}\,|\tilde{\eta}_n|^2 \;\leq\; d^{\frac{1}{2}} \sum_{k=1}^{n} \frac{a^2}{k}\, \mathbb{E}\,|A_{n,k}|_{\mathrm{Op}}^2\, |\Sigma_k|_{\mathrm{F}} \;\leq\; \left(d^{\frac{1}{2}}\, a^2\, C_A^2\right) \frac{1}{n^{2\nu}} \sum_{k=1}^{n} k^{2\nu-1}\, \mathbb{E}\,|\Sigma_k|_{\mathrm{F}} \tag{A7b}$$

where d is the dimension of the parameter space Θ, and $|\Sigma_k|_{\mathrm{F}}$ denotes the Frobenius matrix norm. However, it follows from (u1) that there exists a uniform upper bound S on $|\Sigma_k|_{\mathrm{F}}$. Therefore, by (A7b)

$$\mathbb{E}\,|\tilde{\eta}_n|^2 \;\leq\; \left(d^{\frac{1}{2}}\, a^2\, C_A^2\right) \frac{S}{n^{2\nu}} \sum_{k=1}^{n} k^{2\nu-1} \tag{A7c}$$

Since $\nu > 0$, the right-hand side of (A7c) remains bounded as $n \to \infty$, by the Euler-Maclaurin formula [37]. This immediately yields Condition (27b). □

Condition (27c): to verify this condition, it is first admitted that the following limit is known to hold

$$\lim\, \mathbb{E}\,(\Sigma_k) \;=\; \Sigma^* \tag{A8a}$$

where Σ^* was defined in (4a). Then, let the sum in (27c) be written

$$\sum_{k=1}^{n} \frac{a^2}{k}\, A_{n,k} \Sigma_k A_{n,k} \;=\; \sum_{k=1}^{n} \frac{a^2}{k}\, A_{n,k} \Sigma^* A_{n,k} \;+\; \sum_{k=1}^{n} \frac{a^2}{k}\, A_{n,k}\, [\,\Sigma_k - \Sigma^*]\, A_{n,k} \tag{A8b}$$

Due to the equivalence $A_{n,k} \sim \exp(\ln(n/k) A)$ (see [10], Page 125), the first term in (A8b) is a Riemann sum for the integral [10,34]

$$a^2 \int_0^1 e^{-\ln(s)\, A}\, \Sigma^*\, e^{-\ln(s)\, A}\, d\ln(s) \;=\; a^2 \int_0^{\infty} e^{-t\, A}\, \Sigma^*\, e^{-t\, A}\, dt$$

which is known to be the solution Σ of Lyapunov's equation (7). The second term in (A8b) can be shown to converge to zero in probability, using inequality (A5) and the limit (A8a), by a similar argument to the ones in the verification of Conditions (27a) and (27b). Then, Condition (27c) follows immediately. □

Proof of (A8a): recall that $w_k = u_k + v_{k-1}$ where $u_k = u(\theta_{k-1}, x_k)$ and $v_{k-1} = \nabla D(\theta_{k-1})$. Since (w_k) is a sequence of square-integrable martingale differences, it is possible to write, in the notation of (28),

$$\Sigma_k \;=\; \mathbb{E}\left[\, u_k u_k^{\dagger} \,\middle|\, \mathcal{X}_{k-1} \right] \;-\; v_{k-1} v_{k-1}^{\dagger} \tag{A9a}$$

By (18b), the second term in (A9a) converges to zero almost surely, as $k \to \infty$. It also converges to zero in expectation, since $\nabla D(\theta)$ is uniformly bounded for θ in the compact set Θ^*. For the first term in (A9a), since the x_k are i.i.d. with distribution P_{θ^*}, it follows that

$$\mathbb{E}\left[u_k u_k^\dagger \,\middle|\, \mathcal{X}_{k-1}\right] \;=\; E_{\theta^*}\left[u(\theta_{k-1}, x) u^\dagger(\theta_{k-1}, x)\right] \tag{A9b}$$

Since $u(\theta, x)$ is a continuous vector field on Θ for each $x \in X$, and θ_{k-1} converge to θ^* almost surely, it follows that $u(\theta_{k-1}, x)$ converge to $u(\theta^*, x)$ for each $x \in X$, almost surely. On the other hand, it follows from (u2) that the functions under the expectation E_{θ^*} in (A9b) have bounded second order moments, so they are uniformly integrable [20]. Therefore,

$$\lim\; E_{\theta^*}\left[u(\theta_{k-1}, x) u^\dagger(\theta_{k-1}, x)\right] \;=\; E_{\theta^*}\left[u(\theta^*, x) u^\dagger(\theta^*, x)\right] \;=\; \Sigma^* \tag{A9c}$$

almost surely, by the definition (4a) of Σ^*. It now follows from (A9a), (A9b), and (A9c) that the following limit holds

$$\lim\; \Sigma_k \;=\; \Sigma^* \qquad \text{almost surely} \tag{A9d}$$

To obtain (A8a) it is enough to note, as already stated in the verification of Condition (27b), that the Σ_k are uniformly bounded in the Frobenius matrix norm. Thus, (A9d) implies (A8a), by the dominated convergence theorem. □

Appendix C. Background on the Information Metric

Let $D(\theta)$ be the Kullback–Leibler divergence (2a) or any other so-called α-divergence [19]. Assume the Riemannian metric $\langle \cdot, \cdot \rangle$ of Θ coincides with the information metric of the model P. Then, for any local coordinates $(\tau^\alpha; \alpha = 1, \ldots, d)$, with origin at θ^*, the following relation holds, by definition of the information metric (see [19], Page 54),

$$\left.\frac{\partial^2 D}{\partial \tau^\alpha \partial \tau^\beta}\right|_{\tau^\alpha = 0} \;=\; \left\langle \frac{\partial}{\partial \tau^\alpha}, \frac{\partial}{\partial \tau^\beta} \right\rangle_{\theta^*} \tag{A10}$$

where $\frac{\partial}{\partial \tau^\alpha}$ denote the coordinate vector fields of the local coordinates τ^α. It is also possible to express (A10) in terms of the Riemannian distance $d(\cdot, \cdot)$, induced by the information metric $\langle \cdot, \cdot \rangle$. Precisely,

$$D(\theta) \;=\; \frac{1}{2}\, d^2(\theta, \theta^*) \;+\; o\left(d^2(\theta, \theta^*)\right) \tag{A11}$$

This follows immediately from the second-order Taylor expansion of $D(\theta)$, since θ^* is a minimum of $D(\theta)$, by using (A10). Formula (A11) shows that the divergence $D(\theta)$ is equivalent to half the squared Riemannian distance $d^2(\theta, \theta^*)$, at $\theta = \theta^*$.

The scalar products appearing in (A10) form the components of the information matrix I^τ of the coordinates τ^α,

$$I^\tau_{\alpha\beta} \;=\; \left.\frac{\partial^2 D}{\partial \tau^\alpha \partial \tau^\beta}\right|_{\tau^\alpha = 0}$$

In any change of coordinates, these transform like the components of a $(0, 2)$ covariant tensor [16]. That is, if $(\theta^\alpha; \alpha = 1, \ldots, d)$ are any local coordinates defined at θ^*,

$$I^\tau_{\alpha\beta} \;=\; \left(\frac{\partial \theta^\gamma}{\partial \tau^\alpha}\right)_{\theta^*} I^\theta_{\gamma\kappa} \left(\frac{\partial \theta^\kappa}{\partial \tau^\beta}\right)_{\theta^*}$$

where the subscript θ^* indicates the derivative is evaluated at θ^*, and where $I^\theta_{\gamma\kappa}$ are the components of the information matrix I^θ of the coordinates θ^α.

The recursive estimates θ_n are said to be asymptotically efficient, if they are asymptotically efficient in any local coordinates τ^α, with origin at θ^*. That is, according to the classical definition of asymptotic efficiency [13,14], if the following weak limit of probability distributions is verified [20],

$$\mathcal{L}\left\{(n^{1/2}\tau_n^\alpha)\right\} \implies N_d\left(0,\Sigma^\tau\right) \qquad \Sigma^\tau = \left(I^\tau\right)^{-1} \tag{A12}$$

where $\mathcal{L}\{\ldots\}$ denotes the probability distribution of the quantity in braces, $\tau_n^\alpha = \tau^\alpha(\theta_n)$ are the coordinates of the recursive estimates θ_n, and $N_d\left(0,\Sigma^\tau\right)$ denotes a centred d-variate normal distribution with covariance matrix Σ^τ.

It is important to note that asymptotic efficiency of the recursive estimates θ_n is an intrinsic geometric property, which does not depend on the particular choice of local coordinates τ^α, with origin at θ^*. This can be seen from the transformation rule of the components of the information matrix, described above. In fact, since these transform like the components of a $(0,2)$ covariant tensor, the components of Σ^τ transform like those of a $(2,0)$ contravariant tensor, which is the correct transformation rule for the components of a covariance matrix.

References

1. Bonnabel, S. Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. Automat. Contr.* **2013**, *58*, 2217–2229. [CrossRef]
2. Zhang, H.; Sra, S. First-order methods for geodesically convex optimization. In Proceedings of the 29th Conference on Learning Theory, New York, NY, USA, 23–26 June 2016; pp. 1617–1638.
3. Tripuraneni, N.; Flammarion, N.; Bach, F.; Jordan, M.I. Averaging stochastic gradient descent on Riemannian manifolds. *arXiv* **2018**, arXiv: 1802.09128.
4. Arnaudon, M.; Barbaresco, F.; Yang, L. Riemannian medians and means with applications to radar signal processing. *IEEE J. Sel. Topics Signal Process.* **2013**, *7*, 595–604. [CrossRef]
5. Arnaudon, M.; Dombry, C.; Phan, A.; Yang, L. Stochastic algorithms for computing means of probability measures. *Stochastic Process. Appl.* **2012**, *122*, 1437–1455. [CrossRef]
6. Zhang, H.; Reddi, S.J.; Sra, S. Riemannian SVRG: Fast Stochastic Optimization on Riemannian Manifolds. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 4592–4600.
7. Sun, J.; Qu, Q.; Wright, J. Complete dictionary recovery over the sphere II: recovery by Riemannian trust-region method. *IEEE Trans. Inf. Theory* **2017**, *63*, 885–914. [CrossRef]
8. Ge, R.; Huang, F.; Jin, C.; Yuan, Y. Escaping from saddle points: online stochastic gradient for tensor decomposition. In Proceedings of the Conference on Learning Theory, Paris, France, 3–6 July 2015; pp. 797–842.
9. Duflo, M. *Random Iterative Models*; Springer: Berlin/Heidelberg, Germany, 1997.
10. Nevelson, M.B.; Khasminskii, R.Z.; Khasminskii, R.Z. *Stochastic Approximation and Recursive Estimation*; American Mathematical Society: Providence, RI, USA, 1973.
11. Broniatowski, M. Minimum divergence estimators, maximum likelihood and exponential families. *Stat. Probab. Lett.* **2014**, *93*, 27–33. [CrossRef]
12. Broniatowski, M.; Keziou, A. Parametric estimation and tests through divergences and the duality technique. *J. Multivar. Anal.* **2009**, *100*, 16–36. [CrossRef]
13. Ibragimov, I.A.; Has' Minskii, R.Z. *Statistical Estimation: Asymptotic Theory*; Springer: New York, NY, USA, 1981.
14. Van der Vaart, A.W. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.
15. Benveniste, A.; Métivier, M.; Priouret, P. *Adaptive Algorithms and Stochastic Approximations*; Springer: Berlin/Heidelberg, Germany, 1990.
16. Petersen, P. *Riemannian Geometry*; Springer: New York, NY, USA, 2006.
17. Kushner, H.; Yin, G.G. *Stochastic Approximation and Recursive Algorithms and Applications*; Springer: New York, NY, USA, 2003.
18. Heyde, C.C. *Quasi-Likelihood and Its Applications: A General Approach to Optimal Parameter Estimation*; Springer: New York, NY, USA, 1997.

19. Amari, S.I.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000.
20. Shiryaev, A.N. *Probability-2*; Springer: New York, NY, USA, 1996.
21. Munier, J. Steepest descent method on a Riemannian Manifold: The Convex Case. *Balk. J. Geom. Appl.* **2007**, *12*, 2. Available online: https://hal.archives-ouvertes.fr/hal-00018758 (accessed on 18 October 2019).
22. Udriste, C. *Convex Functions and Optimization Methods on Riemannian Manifolds*; Springer: Berlin/Heidelberg, Germany, 1994.
23. Absil, P.A.; Mahony, R.; Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*; Princeton University Press: Princeton, NJ, USA, 2008.
24. Fang, K.T.; Kotz, S. *Symmetric Multivariate and Related Distributions*; Chapman and Hall: London, UK, 1990.
25. Sra, S.; Hosseini, R. Conic geometric optimization on the manifold of positive definite matrices. *SIAM J. Opt.* **2015**, *25*, 713–739. [CrossRef]
26. Pascal, F.; Bombrun, L.; Tourneret, J.Y.; Berthoumieu, Y. Parameter estimation for multivariate generalized Gaussian distributions. *IEEE Trans. Signal Process.* **2013**, *61*, 5960–5971. [CrossRef]
27. Berkane, M.; Oden, K.; Bentler, P.M. Geodesic estimation in elliptical distributions. *J. Multivar. Anal.* **1997**,*63*, 35–46. [CrossRef]
28. Mostajeran, C.; Sepulchre, R. Sepulchre, Ordering positive definite matrices. *Inf. Geom.* **2018**, *1*, 287–313. [CrossRef]
29. Pennec, X.; Fillard, P.; Ayache, N. A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **2006**, *66*, 41–66. [CrossRef]
30. Higham, N.J. *Functions of Matrices: Theory and Computation*; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2008.
31. Andrieu, C.; Moulines, É.; Priouret, P. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Opt.* **2006**, *44*, 283–312. [CrossRef]
32. Ghadimi, S.; Lan, G. Stochastic first- and zeroth order methods for nonconvex stochastic programming. *SIAM J. Opt.* **2013**, *23*, 2341–2368. [CrossRef]
33. Bottou, L. *On-Line Learning in Neural Networks*; Cambridge University Press: Cambridge, UK, 1998; pp. 9–42.
34. Kailath, T. *Linear Systems*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1980.
35. Heyde, C.C. *Martingale Limit Theory and Its Applications*; Academic Press: Cambridge, MA, USA, 1980.
36. Chavel, I. *Riemannian Geometry, a Modern Introduction*; Cambridge University Press: Cambridge, UK, 2006.
37. Courant, R.; John, F. *Introduction to Calculus and Analysis*; Interscience Publishers: New York, NY, USA, 1965; Volume 1.