



HAL
open science

Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation

Stéphanie Allasonniere, Estelle Kuhn

► **To cite this version:**

Stéphanie Allasonniere, Estelle Kuhn. Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation. 2013. hal-00720617v4

HAL Id: hal-00720617

<https://hal.science/hal-00720617v4>

Preprint submitted on 6 Sep 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation

Stéphanie Allasonnière · Estelle Kuhn

Received: date / Accepted: date

Abstract Estimation in the deformable template model is a big challenge in image analysis. The issue is to estimate an atlas of a population. This atlas contains a template and the corresponding geometrical variability of the observed shapes. The goal is to propose an accurate algorithm with low computational cost and with theoretical guaranties of relevance. This becomes very demanding when dealing with high dimensional data which is particularly the case of medical images. We propose to use an optimized Monte Carlo Markov Chain method into a stochastic Expectation Maximization algorithm in order to estimate the model parameters by maximizing the likelihood. In this paper, we present a new Anisotropic Metropolis Adjusted Langevin Algorithm which we use as transition in the MCMC method. We first prove that this new sampler leads to a geometrically uniformly ergodic Markov chain. We prove also that under mild conditions, the estimated parameters converge almost surely and are asymptotically Gaussian distributed. The methodology developed is then tested on handwritten digits and some 2D and 3D medical images for the deformable model estimation. More widely, the proposed algorithm can be used for a large range

of models in many fields of applications such as pharmacology or genetic.

Keywords Deformable template · geometric variability · maximum likelihood estimation · missing variable · high dimension · stochastic EM algorithm · MCMC · Anisotropic MALA

1 Introduction

We consider here the deformable template model introduced for Computational Anatomy in [18]. This model, which has demonstrated great impact in image analysis, was developed and analyzed later on by many groups (among other [28, 24, 33, 27]). It offers several major advantages. First, it enables to describe the population of interest by a digital anatomical template. It also captures the geometric variability of the population shapes through the modeling of deformations of the template which match it to the observations. Moreover, the metric on the space of deformations is specified in the model as a quantification of the deformation cost. Not only describing the population, this generative model also allows to sample synthetic data using both the template and the geometrical metric of the deformation space which together define the atlas. Nevertheless, the key statistical issue is how to estimate efficiently and accurately these parameters of the model from an observed population of images.

Several numerical methods have been developed mainly for the estimation of the template image (for example [11, 20]). Even if these methods lead to visual interesting results on

S. Allasonnière
CMAP Ecole Polytechnique
Route de Saclay
91128 Palaiseau, FRANCE Tel.: +331.69.33.45.65
E-mail: stephanie.allasonniere@polytechnique.edu

E. Kuhn
INRA
Domaine de Vilvert
78352 Jouy-en-Josas, FRANCE

some training samples, they suffer from a lack of theoretical properties raising the question of the relevance of the output and are not robust to noisy data. Another important contribution toward the statistical formulation of the template estimation issue was proposed in [17]. However interesting this approach is not entirely satisfactory since the deformations are applied to discrete observations requiring some interpolation. Moreover it does not formulate the analysis in terms of a generative model which appears very attractive as mentioned above. To overcome these lacks, a coherent statistical generative model was formulated in [2]. For estimating all the model parameters, the template image together with the geometrical metric, the authors proposed a deterministic algorithm based on an approximation of the well-known Expectation Maximization (EM) algorithm (see [14]), where the posterior distribution is replaced by a Dirac measure on its mode (called FAM-EM). However, such an approximation leads to the non-convergence of the estimates highlighted when considering noisy observations.

One solution to face this problem is to consider a convergent stochastic approximation of the EM (SAEM) algorithm which was proposed in [13]. An extension using Monte Carlo Markov Chain (MCMC) methods was developed and studied in [21] and [5] allowing for wider applications. To apply this extension to the deformable template model, the authors in [5] chose a Metropolis Hastings within Gibbs sampler (also called hybrid Gibbs) as MCMC method since the variables to sample were of large dimension (the usual Metropolis Hastings algorithm providing low acceptance rates). This estimation algorithm has been proved convergent and performs very well on very different kind of data as presented in [4]. Nevertheless, the hybrid Gibbs sampler becomes computationally very expensive when sampling very high dimensional variables. Although it reduces the dimension of the sampling to one which enables to stride easier the target density support, it loops over the sampling variable coordinates, which becomes computationally unusable as soon as the dimension is very large or as the acceptance ratio involves heavy computations. To overcome the problem of computational cost of this estimation algorithm, some authors propose to simplify the statistical model constrain-

ing the correlations of the deformations (see [29, 22]).

Our purpose in this paper is to propose an efficient and convergent estimation algorithm for the deformable template model in high dimension without any constraints. With regards to the above considerations, the computational cost of the estimation algorithm can be reduced by optimizing the sampling scheme in the MCMC method.

The sampling of high dimensional variables is a well-known difficult challenge. In particular, many authors have proposed to use the Metropolis Adjusted Langevin Algorithm (MALA) (see [30] and [31]). This algorithm is a particular random walk Metropolis Hastings sampler. Starting from the current iterate of the Markov chain, one simulates a candidate with respect to a Gaussian proposal with an expectation equal to the sum of this current iterate and a drift related to the target distribution. The covariance matrix is diagonal and isotropic. This candidate is accepted or rejected with a probability given by the Metropolis Hastings acceptance ratio.

Some modifications have been proposed in particular to optimize the covariance matrix of the proposal in order to better stride the support of the target distribution (see [32, 7, 23, 16]). In [7] and [23], the authors proposed to construct adaptive MALA chains for which they prove the geometric ergodicity of the chain uniformly on any compact subset of its parameters. Unfortunately, this technique does not take the whole advantage of changing the proposal using the target distribution. In particular, the covariance matrix of the proposal is given by a stochastic approximation of the empirical covariance matrix. This choice seems completely relevant as soon as the convergence toward the stationary distribution is reached. However, it does not provide a good guess of the variability during the first iterations of the chain since it is still very dependent on the initialization. This leads to chains that may be numerically trapped. Moreover, this particular algorithm may require a lot of tuning parameters. Although the theoretical convergence is proved, this algorithm may be very difficult to optimize in practice *into* an estimation process.

Recently, the authors in [16] proposed the Riemann manifold Langevin algorithm in order to sample from a target density in high dimen-

sional setting with strong correlations. This algorithm is also a MALA based one for which the choice of the proposal covariance is guided by the metric of the underlying Riemann manifold. It requires to evaluate the metric, its inverse as well as its derivatives. The proposed well-suited metric is the Fisher-Rao information matrix or its empirical value. However, in the context we are dealing with, the real metric, namely the metric of the space of non-rigid deformations, is not explicit preventing from any use of it (the simplest case of the 3-landmark-matching problem is calculated in [26] leading to a very intricate formula which is difficult to extend to more complex models). Moreover, if we consider the constant curvature simplification suggested in [16], one still needs to invert the metric which may be neither explicit nor computationally tractable. Note that these constraints are common with other application fields such as genetic or pharmacology, where models are often complex.

For all these reasons, we propose to adapt the MALA algorithm in the spirit of both works in [7] and [16] to get an efficient sampler into the stochastic EM algorithm. Therefore, we propose to sample from a proposal distribution which has the same expectation as the MALA but using a full anisotropic covariance matrix based on the anisotropy and correlations of the target distribution. This sampler will be called AMALA in the sequel. The expectation is obtained as the sum of the current iterate plus a drift which is proportional to the gradient of the logarithm of the target distribution. We construct the covariance matrix as a regularization of the Gram matrix of this drift. We prove the geometric ergodicity uniformly on any compact set of the AMALA assuming some regularity conditions on the target distribution. We also prove the almost sure convergence of the parameter estimated sequence generated by the coupling of AMALA and SAEM algorithms (AMALA-SAEM) toward the maximum likelihood estimate under some regularity assumptions on the model. Moreover, we prove a Central Limit Theorem for this sequence under usual conditions on the model.

We test our estimation algorithm on the deformable template model for estimating handwritten digit atlases from the USPS database and medical images of corpus callosum (2D) and of dendrite spine excrescences (3D). The

proposed estimation method is compared with the results obtained from the FAM-EM algorithm and from the MCMC-SAEM algorithm using different samplers namely the hybrid Gibbs sampler, the MALA and the adaptive MALA proposed in [7] previously introduced. The comparison is also made via classification rates on the USPS database. These experiments demonstrate the good behavior of our method in both the accuracy of the estimation and the low computational cost in high dimension.

The paper is organized as follows. In Section 2, we recall the Bayesian Mixed Effect (BME) template model. In Section 3, we consider the maximum likelihood estimation issue in the general framework of missing data models. We present our stochastic version of the EM algorithm using the AMALA sampler. The convergence properties are established in Section 4. Section 5 is devoted to the experiments on the BME template estimation. Finally, we give some conclusion in Section 6. The proofs are postponed in Section 7.

2 Description of the Bayesian Mixed Effect (BME) Template model

The deformable template model aims at summarizing a population of images by two quantities. The first one is a mean image called template which has to represent a relevant shape as one could find in the population. The second quantity represents the variance in the space of shapes. This corresponds to the geometrical variability around the mean shape. Let us now describe the deformable template model more precisely.

We consider the hierarchical Bayesian framework for dense deformable template developed in [2] where each image in a population is assumed to be generated as a noisy and randomly deformed version of the template.

The database is composed of n grey level images $(y_i)_{1 \leq i \leq n}$ observed on a grid A of pixels (or voxels) included in a continuous domain $D \subset \mathbb{R}^d$, (typically $D = [-1, 1]^d$ where d equals 2 or 3). The expected template $I_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ takes its values in the continuous domain. Each observation y is assumed to be a discretization on A of a random deformation of this template plus an independent noise. Therefore, there exists an unobserved deformation field (also called

mapping) $m : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that for $u \in \Lambda$

$$y(u) = I_0(v_u - m(v_u)) + \sigma\epsilon(u),$$

where $\sigma\epsilon$ denotes the independent additive noise and v_u is the location of pixel (or voxel) u .

Considering the template and the deformations as continuous functions would lead to a dense problem. The dimension is reduced assuming that both elements belong to a subset of fixed Reproducing Kernel Hilbert Spaces (RKHS) V_p and V_g defined by their respective kernels K_p and K_g . More precisely, let $(r_{p,j})_{1 \leq j \leq k_p}$ -respectively $(r_{g,j})_{1 \leq j \leq k_g}$ - be some fixed control points in the domain D : there exist $\alpha \in \mathbb{R}^{k_p}$ -resp. $z \in \mathbb{R}^{k_g} \times \mathbb{R}^{k_g}$ - such that for all v in D :

$$I_\alpha(v) = (\mathbf{K}_p \alpha)(v) = \sum_{j=1}^{k_p} K_p(v, r_{p,j}) \alpha^j \quad (1)$$

$$m_z(v) = (\mathbf{K}_g z)(v) = \sum_{j=1}^{k_g} K_g(v, r_{g,j}) z^j. \quad (2)$$

For clarity, we write $\mathbf{y} = (y_i)_{1 \leq i \leq n}$ for the n -tuple of observations and $\mathbf{z} = (z_i)_{1 \leq i \leq n}$ for the n -tuple of unobserved variables defining the deformations. The statistical model on the observations is chosen as follows:

$$\begin{cases} \mathbf{z} \sim \otimes_{i=1}^n \mathcal{N}_{dk_g}(0, \Gamma_g) \mid \Gamma_g, \\ \mathbf{y} \sim \otimes_{i=1}^n \mathcal{N}_{|\Lambda|}(m_{z_i} I_\alpha, \sigma^2 Id_{|\Lambda|}) \mid \mathbf{z}, \alpha, \sigma^2, \end{cases} \quad (3)$$

where \otimes denotes the product of independent variables and $m I_\alpha(u) = I_\alpha(v_u - m(v_u))$, for u in Λ . The parameters of interest are the template α , the noise variance σ^2 and the deformation covariance matrix Γ_g . We assume that $\theta = (\alpha, \sigma^2, \Gamma_g)$ belongs to the parameter space Θ :

$$\Theta \triangleq \{ \theta = (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p}, |\alpha| < R, \sigma > 0, \Gamma_g \in \text{Sym}_{dk_g, *}^+(\mathbb{R}) \}, \quad (4)$$

where $\text{Sym}_{dk_g, *}^+(\mathbb{R})$ is the cone of real positive $dk_g \times dk_g$ definite symmetric matrices, R is an arbitrary positive constant and d is the space dimension (typically 2 or 3 for images).

Since we aim at dealing with small size samples and high dimensional parameters, we work in the Bayesian framework and we introduce priors on the parameters. In addition of guiding the estimation it regularizes the estimation

as shown in [2]. The priors are all independent: $\theta = (\alpha, \sigma^2, \Gamma_g) \sim \nu_p \otimes \nu_g$ where

$$\begin{cases} \nu_p(d\alpha, d\sigma^2) \propto \exp\left(-\frac{1}{2}(\alpha - \mu_p)^T (\Sigma_p)^{-1} (\alpha - \mu_p)\right) \times \\ \left(\exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right) \frac{1}{\sqrt{\sigma^2}}\right)^{a_p} d\sigma^2 d\alpha, \quad a_p \geq 3, \\ \nu_g(d\Gamma_g) \propto \left(\exp(-\langle \Gamma_g^{-1}, \Sigma_g \rangle_F / 2) \frac{1}{\sqrt{|\Gamma_g|}}\right)^{a_g} d\Gamma_g, \\ a_g \geq 4k_g + 1. \end{cases} \quad (5)$$

For two matrices A, B we define the Frobenius inner product by $\langle A, B \rangle_F \triangleq \text{tr}(A^T B)$.

Parameter estimation for this model is then performed by Maximum A Posteriori (MAP) :

$$\hat{\theta} = \underset{\theta \in \Theta}{\text{argmax}} q(\theta | \mathbf{y}), \quad (6)$$

where $q(\theta | \mathbf{y})$ is the posterior density of θ conditional on \mathbf{y} . The existence and consistency of the MAP estimator for the BME template model has been proved in [2].

Note that this model belongs to a more general class called mixed effect models. The fixed effects are the parameters θ and the random effects are the deformation coefficients \mathbf{z} . The estimation issue in this class is treated in the same way as the likelihood maximization problem in the more general framework of incomplete-data models. Therefore, the next section will be presented in this general setting in which the proposed algorithm applies.

3 Maximum likelihood estimation

3.1 Maximum likelihood estimation for incomplete data setting

We consider in this section the standard incomplete data (or partially-observed-data) setting and recall the usual notation. We denote by $y \in \mathbb{R}^q$ the observed data and by $z \in \mathbb{R}^l$ the missing data, so that we obtain the complete data $(y, z) \in \mathbb{R}^{q+l}$ for some $q \in \mathbb{N}^*$ and $l \in \mathbb{N}^*$. We consider these data as random vectors. Let μ' be a σ -finite measure on \mathbb{R}^{q+l} and μ the restriction of μ' to \mathbb{R}^l generated by the projection $(y, z) \mapsto z$. We assume that the probability density function (pdf) of the random vector (y, z) belongs to $\mathcal{P} = \{f(y, z; \theta), \theta \in \Theta\}$, a family of parametric probability density functions on

\mathbb{R}^{q+l} w.r.t. μ' , where $\Theta \subset \mathbb{R}^p$. Therefore, the observed likelihood (i.e. the incomplete-data likelihood) is defined for some $\theta \in \Theta$ by:

$$g(y; \theta) \triangleq \int f(y, z; \theta) \mu(dz). \quad (7)$$

Our purpose is to find the maximum likelihood estimate that is the value $\hat{\theta}_g$ in Θ that maximizes the observed likelihood g given a sample of observations. However, this maximization can often not be done analytically because of the integration involved in (7). A powerful tool which enables to compute this maximization in such a setting is the Expectation Maximization (EM) algorithm (see [14]). It is an iterative procedure which consists of two steps. First, the so-called E-step computes the conditional expectation of the complete log-likelihood using the current parameter value. Second, the M-step achieves the update of the parameter by maximizing this expectation over Θ . However, the computation of this expectation is often intractable analytically. Therefore, alternative procedures have been proposed. We are particularly interested in the Stochastic Approximation EM (SAEM) algorithm (see [13]) because of its theoretical convergence property and its small computation time. In this stochastic algorithm, the usual E-step is replaced by two steps, the first one corresponding to the simulation of realizations of the missing data, the second one to the computation of a stochastic approximation of the complete log-likelihood using these simulated values. It can be shown under weak regularity conditions that the sequence generated by this algorithm converges almost surely toward a local maximum of the observed likelihood (see [13]).

Nevertheless the simulation step requires some attention. In the SAEM algorithm the simulated values of the missing data have to be drawn from the posterior distribution defined by:

$$p(z|y; \theta) \triangleq \begin{cases} f(y, z; \theta)/g(y; \theta) & \text{if } g(y; \theta) \neq 0 \\ 0 & \text{if } g(y; \theta) = 0. \end{cases}$$

When not possible, the extension using MCMC method (see [21, 5]) allows to apply the SAEM algorithm using simulations obtained from some transition probability of an ergodic Markov chain having the targeted posterior distribution as stationary distribution. Methods like Metropolis Hastings algorithm or Gibbs sampler are useful to perform this assignment. However, this

becomes very challenging in high dimensional setting. Indeed, when the MCMC procedure has to explore a space of high dimension, its convergence may occur in practice only after a possibly infinite time. Thus, it is necessary to optimize this MCMC procedure. This is what we will propose in the following paragraph.

3.2 Description of the sampling method: Anisotropic Metropolis Adjusted Langevin Algorithm

We propose an anisotropic version of the well-known Metropolis Adjusted Langevin Algorithm (MALA). So let us first recall the steps of this algorithm. Let \mathcal{X} be an open subset of \mathbb{R}^l , the l -dimensional Euclidean space equipped with its Borel σ -algebra \mathcal{B} . Let us denote π the pdf of the target distribution with respect to the Lebesgue measure on \mathcal{X} . We assume that π is positive continuously differentiable. At each iteration k of this algorithm, a candidate X_c is simulated with respect to the Gaussian distribution with expectation $X_k + \frac{\sigma^2}{2} D(X_k)$ and covariance $\sigma^2 Id_l$ where X_k is the current value,

$$D(x) = \frac{b}{\max(b, |\nabla \log \pi(x)|)} \nabla \log \pi(x), \quad (8)$$

Id_l is the identity matrix in \mathbb{R}^l and $b > 0$ is a fixed truncation threshold. Note that the truncation of the drift D was already suggested in [15] to provide more stability. In the following, we denote $q_{MALA}(x, \cdot)$ the pdf of this Gaussian candidate distribution starting from x . Given this candidate, the next value of the Markov chain is updated using an acceptance ratio $\alpha_{MALA}(X_k, X_c)$ as follows: $X_{k+1} = X_c$ with probability

$$\alpha_{MALA}(X_k, X_c) = \min \left(1, \frac{\pi(X_c) q_{MALA}(X_c, X_k)}{q_{MALA}(X_k, X_c) \pi(X_k)} \right) \quad (9)$$

and $X_{k+1} = X_k$ with probability $1 - \alpha_{MALA}(X_k, X_c)$. This provides a transition kernel Π_{MALA} of this form: for any Borel set $A \in \mathcal{B}$

$$\begin{aligned} \Pi_{MALA}(x, A) &= \int_A \alpha_{MALA}(x, z) q_{MALA}(x, z) dz + \\ & \mathbb{1}_A(x) \int_{\mathcal{X}} (1 - \alpha_{MALA}(x, z)) q_{MALA}(x, z) dz. \end{aligned} \quad (10)$$

The Gaussian proposal of the MALA algorithm is optimized with respect to its expectation guided by the Langevin diffusion. One step further is to optimize also its covariance matrix. A first work in this direction was proposed in [7]. The covariance matrix of the proposal is given by a projection of a stochastic approximation of the empirical covariance matrix. It produces an adaptive Markov chain. This process involves some additional tuning parameters which have to be calibrated. Since our goal is to use this sampler in an estimation algorithm, the sampler has at each iteration a different target distribution (depending on the current estimate of the parameter). Therefore, the optimal tuning parameter may be different along the iterations of the estimation process. Although we agree with the idea of using adaptive chain, we prefer taking the advantage of the dynamic of the estimation algorithm. On the other side, an intrinsic solution has been proposed in [16] where the covariance matrix is given by the metric of the Riemann manifold of the variable to sample. Unfortunately, this metric may not be accessible and its empirical approximation not easy to compute. This is particularly the case in the BME template model.

For these reasons, we propose a sampler in the spirit of [7], [16] or [23] however not providing an adaptive chain as motivated above. The adaption comes from the dependency of the target distribution with respect to the parameters of the model which are updated along the estimation algorithm. The proposal remains a Gaussian distribution but both the drift and the covariance matrix depend on the gradient of the target distribution. At the k^{th} iteration, we are provided with X_k . The candidate is sampled from the Gaussian distribution with expectation $X_k + \delta D(X_k)$ and covariance matrix $\delta \Sigma(X_k)$ denoted in the sequel $\mathcal{N}(X_k + \delta D(X_k), \delta \Sigma(X_k))$ where $\Sigma(x)$ is given by :

$$\Sigma(x) = \varepsilon Id_l + D(x)D(x)^T, \quad (11)$$

D is defined in Equation (8) and $\varepsilon > 0$ is a small regularization parameter. Note that the threshold parameter b leads to a symmetric positive definite covariance matrix with bounded non zero eigenvalues. We introduce the gradient of

$\log \pi$ into the covariance matrix to provide an anisotropic covariance matrix depending on the amplitude of the drift at the current value. When the drift is large, the candidate is likely to be far from the current value. This large step may not be of the right amplitude and a large variance will enable more flexibility. Moreover, this enables to explore a larger area around these candidates which would not be possible with a fixed variance. On the other hand, when the drift is small in a particular direction, it means that the current value is within a region of high probability for the next value of the Markov chain. Therefore, the candidate should not move too far neither with a large drift nor with a large variance. This enables to sample a lot around large modes which is of particular interest. This covariance also enables to treat the directions of interest with different amplitudes of variances as the drift already does. It also provides dependencies between coordinates since the directions of large variances are likely to be different from the Euclidean axis. This is taken into account here by introducing the Gram matrix of the drift into the covariance matrix.

We denote by q_c the pdf of this proposal distribution. The transition kernel becomes: for any Borel set A .

$$\begin{aligned} \Pi(x, A) = & \int_A \alpha(x, z) q_c(x, z) dz + \\ & \mathbb{1}_A(x) \int_{\mathcal{X}} (1 - \alpha(x, z)) q_c(x, z) dz, \end{aligned} \quad (12)$$

where

$$\alpha(X_k, X_c) = \min \left(1, \frac{\pi(X_c) q_c(X_c, X_k)}{q_c(X_k, X_c) \pi(X_k)} \right). \quad (13)$$

3.3 Description of the stochastic estimation algorithm

Back to the stochastic estimation algorithm, the target distribution of the sampler is the posterior distribution $p(\cdot | y; \theta)$.

The four steps of the proposed AMALA-SAEM algorithm are detailed in this subsection : simulation, stochastic approximation, truncation on random boundaries and maximization steps. At each iteration k of the algorithm, simulated values of the missing data are drawn from the transition probability of the AMALA

algorithm described in Section 3.2 with the current value of the parameters. Then, a stochastic approximation of the complete log-likelihood is computed using these simulated values for the missing data and is truncated using random boundaries. Finally, the parameters are updated by maximizing this quantity over Θ .

We consider here only parametric models \mathcal{P} which belong to the curved exponential family, this means that the complete likelihood $f(y, z; \theta)$ can be written as:

$$f(y, z; \theta) = \exp[-\psi(\theta) + \langle S(z), \phi(\theta) \rangle],$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product, the sufficient statistics S is a function on \mathbb{R}^l , taking its values in a subset \mathcal{S} of \mathbb{R}^m and ψ, ϕ are two functions on Θ (note that S, ϕ and ψ may depend also on y , but we omit this dependency for simplicity). This condition is usual in the framework of EM algorithm applications and it is fulfilled by large range of models even complex ones as the BME template model. Therefore the stochastic approximation can be done either on the sufficient statistics S of the model or on the complete log-likelihood using a positive step-size sequence $(\gamma_k)_{k \in \mathbb{N}}$.

Concerning the truncation procedure, we introduce a sequence of increasing compact subsets of \mathcal{S} denoted by $(\mathcal{K}_q)_{q \geq 0}$ such that $\cup_{q \geq 0} \mathcal{K}_q = \mathcal{S}$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1})$, for all $q \geq 0$. Let also $(\varepsilon_q)_{q \geq 0}$ be a monotone non-increasing sequence of positive numbers and K a compact subset of \mathbb{R}^l . At iteration k we simulate a value \bar{z} for the missing data from the Anisotropic Metropolis Adjusted Langevin Algorithm using the current value of the parameter θ_{k-1} . We compute the associated stochastic approximation of the sufficient statistics of the model \bar{s} . If it does not wander outside the current compact set \mathcal{K}_k and if it is not too far from its previous value s_{k-1} , we keep the possible proposed values for (z_k, s_k) . As soon as one of these conditions is not fulfilled, we reinitialize the sequences of z and s using a projection (for more details see [6]) and we increase the size of the compact set used for the truncation. As explained in [6], the re-projections act as a drift as they force the chain to come back to a compact set when it grows too rapidly. It reinitializes the algorithm with a smaller step size. However, as the chain has an unbounded support, it requires the use of adaptive truncations. As we shall see in the Proof section (and already noted in [6]), the limitation imposed on the increments of the

sequence is required in order to ensure the convergence of the whole algorithm.

Concerning the maximization step, we denote by L the function defined on $\mathcal{S} \times \Theta$ taking values in \mathbb{R} equaled for all (s, θ) to $L(s, \theta) = -\psi(\theta) + \langle s, \phi(\theta) \rangle$. We assume that there exists a function $\hat{\theta}$ defined on \mathcal{S} taking values in Θ such that

$$\forall \theta \in \Theta \forall s \in \mathcal{S} L(s, \hat{\theta}(s)) \geq L(s, \theta).$$

Finally we update the parameter using the value of the function $\hat{\theta}$ evaluated in s_k .

The complete algorithm is summarized in Algorithm 1. It only involves three parameters: b the threshold for the gradient which appears in the expectation as well as in the covariance matrix, δ the scale on this gradient and ε a small regularization parameter to ensure a positive definite covariance matrix. The scale δ can be easily optimized looking at the data we are dealing with to adapt to the range of the drift. The value of the threshold b is in practice never reached. The practical choices for the sequences $(\gamma_k)_k$ and $(\varepsilon_k)_k$ of positive step sizes used in the stochastic approximation and the tuning parameters will be detailed in the section devoted to the experiments.

4 Theoretical Properties

4.1 Geometric ergodicity of the AMALA

Let \mathcal{S} be a subset of \mathbb{R}^m for some positive integer m . Let \mathcal{X} be a measurable subspace of \mathbb{R}^l for some positive integer l . Let $(\pi_s)_{s \in \mathcal{S}}$ be a family of positive continuously differentiable probability density functions with respect to the Lebesgue measure on \mathcal{X} . For any $s \in \mathcal{S}$, denote by Π_s the transition kernel corresponding to the AMALA procedure described in Section 3.2 with stationary distribution π_s . We prove in the following proposition that each kernel of the family $(\Pi_s)_{s \in \mathcal{S}}$ is uniformly geometrically ergodic and that this property holds uniformly in s on any compact subset \mathcal{K} of \mathcal{S} .

We require a usual assumption on the stationary distributions namely the so-called super-exponential property given by:

(B1) For all $s \in \mathcal{S}$, the density π_s is positive with continuous first derivative such that:

$$\lim_{|x| \rightarrow \infty} n(x) \cdot \nabla \log \pi_s(x) = -\infty \quad (14)$$

Algorithm 1 AMALA within SAEM

for all $k = 1 : k_{end}$ **do**

Sample z_c with respect to $\mathcal{N}(z_{k-1} + \delta D(z_{k-1}, \theta_{k-1}), \delta \Sigma(z_{k-1}, \theta_{k-1}))$ whose pdf is denoted $q_{s_{k-1}}(z_{k-1}, \cdot)$ where

$$\left\{ \begin{array}{l} D(z_{k-1}, \theta_{k-1}) = \frac{b \nabla \log p(z_{k-1} | y; \theta_{k-1})}{\max(b, |\nabla \log p(z_{k-1} | y; \theta_{k-1})|)} \\ \Sigma(z_{k-1}, \theta_{k-1}) = D(z_{k-1}, \theta_{k-1}) D(z_{k-1}, \theta_{k-1})^T + \varepsilon Id_l. \end{array} \right.$$

Compute the acceptance ratio $\alpha_{\theta_{k-1}}(z_{k-1}, z_c)$ as defined in Eq. (13).

Sample $\bar{z} = z_c$ with probability $\alpha_{\theta_{k-1}}(z_{k-1}, z_c)$ and $\bar{z} = z_{k-1}$ with probability $1 - \alpha_{\theta_{k-1}}(z_{k-1}, z_c)$

Do the stochastic approximation

$$\bar{s} = s_{k-1} + \gamma_k (S(\bar{z}) - s_{k-1}),$$

 where $(\gamma_k)_k$ is a sequence of positive step sizes.

if $\bar{s} \in \mathcal{K}_{\kappa_{k-1}}$ and $\|\bar{s} - s_{k-1}\| \leq \varepsilon_{\zeta_{k-1}}$ **then**

Set $(z_k, s_k) = (\bar{z}, \bar{s})$ and $\kappa_k = \kappa_{k-1}$, $\nu_k = \nu_{k-1} + 1$, $\zeta_k = \zeta_{k-1} + 1$

else

set $(z_k, s_k) = (\bar{z}, \bar{s}) \in \mathcal{K} \times \mathcal{K}_0$ and $\kappa_k = \kappa_{k-1} + 1$, $\nu_k = 0$, $\zeta_k = \zeta_{k-1} + \Psi(\nu_{k-1})$

 where $\Psi : \mathbb{N} \rightarrow \mathbb{Z}$ is a function such that $\Psi(k) > -k$ for any k

 and (\bar{z}, \bar{s}) is chosen arbitrarily.

end if

Update the parameter

$$\theta_k = \hat{\theta}(s_k)$$

end for

and

$$\limsup_{|x| \rightarrow \infty} n(x) \cdot m_s(x) < 0 \quad (15)$$

where ∇ is the gradient operator in \mathbb{R}^l , $n(x) = \frac{x}{|x|}$ is the unit vector pointing in the direction of x and $m_s(x) = \frac{\nabla \pi_s(x)}{|\nabla \pi_s(x)|}$ is the unit vector in the direction of the gradient of the stationary distribution at point x .

We assume also some regularity properties of the stationary distributions with respect to s .

(B2) For all $x \in \mathcal{X}$, the functions $s \mapsto \pi_s$ and $s \mapsto \nabla_x \log \pi_s$ are continuous on \mathcal{S} .

We now define for some $\beta \in]0, 1[$, $V_s(x) = c_s \pi_s(x)^{-\beta}$ where c_s is a constant so that $V_s(x) \geq 1$ for all $x \in \mathcal{X}$. Let also $V_1(x) = \inf_{s \in \mathcal{S}} V_s(x)$ and

$$V_2(x) = \sup_{s \in \mathcal{S}} V_s(x).$$

Let us assume conditions on V_2 :

(B3) There exists $b_0 > 0$ such that, for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$, $V_2^{b_0}$ is integrable against $\Pi_s(x, \cdot)$ and

$$\limsup_{b \rightarrow 0} \sup_{s \in \mathcal{S}, x \in \mathcal{X}} \Pi_s V_2^b(x) = 1. \quad (16)$$

Proposition 1 Assume **(B1-B3)**. Let \mathcal{C} a compact subset of \mathcal{S} . There exist a function $V \geq 1$, a set $\mathcal{C} \subseteq \mathcal{X}$, a probability measure ν such that $\nu(\mathcal{C}) > 0$ and there exist constants $\lambda \in]0, 1[$, $b \in [0, \infty[$ and $\varepsilon \in]0, 1[$ such that for all $s \in \mathcal{C}$:

$$\Pi_s(x, A) \geq \varepsilon \nu(A) \quad \forall x \in \mathcal{C} \quad \forall A \in \mathcal{B}, \quad (17)$$

$$\Pi_s V(x) \leq \lambda V(x) + b \mathbf{1}_{\mathcal{C}}(x). \quad (18)$$

The proof of Proposition 1 is given in Appendix.

The first equation defines \mathcal{C} as a small set for the transition kernels (Π_s) . Note that both ε and ν can depend on \mathcal{C} . The ν -small set Equation (17) "in one step" also implies the ν -irreducibility of the transition kernels and their aperiodicity (see [25]). The second inequality is a drift condition which states that the transition kernels tend to bring back elements into the small set. As a consequence of these well known drift conditions, the transition kernels (Π_s) are V -uniformly ergodic. Moreover this property holds uniformly in s in any compact subset $\mathcal{K} \subset \mathcal{S}$. That is to say: for any compact $\mathcal{K} \subset \mathcal{S}$, there exist $0 < \rho < 1$ and $0 < c < \infty$ such that for all $n \in \mathbb{N}^*$ and f such that $\|f\|_V = \sup_{x \in \mathcal{X}} \frac{\|f(x)\|}{V(x)} < \infty$:

$$\sup_{s \in \mathcal{K}} \|\Pi_s^n f(\cdot) - \pi_s f\|_V \leq c \rho^n \|f\|_V. \quad (19)$$

Remark 1 The same property holds for any power p of the function V such that $0 < p\beta < 1$. Indeed, the proof follows the same lines as it can be seen in Section 7. This is a property that will appear useful in the sequel to prove some properties of the estimation algorithm.

4.2 Convergence property of the estimated sequence generated by the AMALA-SAEM algorithm

We do the following assumptions on the model which are quite usual in the context of missing data model using EM-like algorithms (see [13], [21]).

For sake of simplicity we denote in the sequel $p_\theta(\cdot)$ instead of $p(\cdot | y; \theta)$ the posterior distribution.

- (M1) The parameter space Θ is an open subset of \mathbb{R}^p . The complete data likelihood function is given by:

$$f(y, z; \theta) = \exp[-\psi(\theta) + \langle S(z), \phi(\theta) \rangle],$$

where S is a Borel function on \mathbb{R}^l taking its values in an open convex subset \mathcal{S} of \mathbb{R}^m . Moreover, the convex hull of $S(\mathbb{R}^l)$ is included in \mathcal{S} , and, for all θ in Θ ,

$$\int \|S(z)\| p_\theta(z) \mu(dz) < \infty.$$

- (M2) The functions ψ and ϕ are twice continuously differentiable on Θ .
- (M3) The function $\bar{s} : \Theta \rightarrow \mathcal{S}$ defined as

$$\bar{s}(\theta) \triangleq \int S(z) p_\theta(z) \mu(dz)$$

is continuously differentiable on Θ .

- (M4) The function $l : \Theta \rightarrow \mathbb{R}$ defined as the observed-data log-likelihood

$$l(\theta) \triangleq \log g(y; \theta) = \log \int f(y, z; \theta) \mu(dz)$$

is continuously differentiable on Θ and

$$\partial_\theta \int f(y, z; \theta) \mu(dz) = \int \partial_\theta f(y, z; \theta) \mu(dz).$$

- (M5) There exists a function $\hat{\theta} : \mathcal{S} \rightarrow \Theta$, such that:

$$\forall s \in \mathcal{S}, \forall \theta \in \Theta, L(s; \hat{\theta}(s)) \geq L(s; \theta).$$

Moreover, the function $\hat{\theta}$ is continuously differentiable on \mathcal{S} .

- (M6) The functions $l : \Theta \rightarrow \mathbb{R}$ and $\hat{\theta} : \mathcal{S} \rightarrow \Theta$ are m times differentiable.

- (M7)

(i) There exists an $M_0 > 0$ such that

$$\left\{ s \in \mathcal{S}, \partial_s l(\hat{\theta}(s)) = 0 \right\} \subset \left\{ s \in \mathcal{S}, -l(\hat{\theta}(s)) < M_0 \right\}.$$

- (ii) For all $M_1 > M_0$, the set $\overline{\text{Conv}(S(\mathbb{R}^l)) \cap \{s \in \mathcal{S}, -l(\hat{\theta}(s)) \leq M_1\}}$ is a compact set of \mathcal{S} .

- (M8) There exists a polynomial function P such that for all $z \in \mathcal{X}$

$$\|S(z)\| \leq P(z).$$

- (B4) For any compact subset \mathcal{K} of \mathcal{S} , there exists a polynomial function Q of the hidden variable such that $\sup_{s \in \mathcal{K}} |\nabla_z \log p_{\hat{\theta}(s)}(z)| \leq Q(z)$.

Moreover a usual additional assumption is required on the step size sequences of the stochastic approximation.

- (A4) The sequences $\gamma = (\gamma_k)_{k \geq 0}$ and $\varepsilon = (\varepsilon_k)_{k \geq 0}$ are non-increasing, positive and satisfy: there exist $0 < a < 1$ and $p \geq 2$ such that $\sum_{k=0}^{\infty} \gamma_k = \infty$, $\lim_{k \rightarrow \infty} \varepsilon_k = 0$ and $\sum_{k=1}^{\infty} \{\gamma_k^2 + \gamma_k \varepsilon_k^a + (\gamma_k \varepsilon_k^{-1})^p\} < \infty$.

Theorem 1 (Convergence Result for the Estimated Sequence generated by Algorithm 1) Assume (M1-M8) and (A4). Assume that the family of posterior density functions $\{p_{\hat{\theta}(s)}, s \in \mathcal{S}\}$ satisfies assumptions (B1-B4).

Let \mathbf{K} be a compact subset of \mathcal{X} and $\mathcal{K}_0 \subset \{s \in \mathcal{S}, -l(\hat{\theta}(s)) < M_0\} \cap \overline{\text{Conv}(S(\mathbb{R}^l))}$ (where M_0 is defined in (M7)). Then, for all $z_0 \in \mathbf{K}$ and $s_0 \in \mathcal{K}_0$, we have $\lim_{k \rightarrow \infty} d(\theta_k, \mathcal{L}) = 0$ a.s. where $(\theta_k)_k$ is the sequence generated by Algorithm 1 and $\mathcal{L} \triangleq \{\theta \in \Theta, \partial_\theta l(\theta) = 0\}$.

The proof is postponed to Appendix 7.2.

4.3 Central Limit Theorem for the estimated sequence generated by the AMALA-SAEM

Theorem 1 ensures that the number of re-initializations of the sequence of stochastic approximation of Algorithm 1 is finite almost surely. We can therefore consider only the non truncated sequence when we are interested in its asymptotic behavior.

Let us write the stochastic approximation procedure :

$$s_k = s_{k-1} + \gamma_k h(s_{k-1}) + \gamma_k \eta_k$$

where $H_s(z) = S(z) - s$, $h(s) = \mathbb{E}_{p_{\hat{\theta}(s)}}(H_s(z))$, $\eta_k = S(z_k) - \mathbb{E}_{p_{\hat{\theta}(s_{k-1})}}(S(z))$ and $\mathbb{E}_{p_{\hat{\theta}(s)}}$ is the expectation under the invariant measure $p_{\hat{\theta}(s)}$.

Let us introduce some usual assumptions in the spirit of these of Delyon (see [12]).

- (N1) The sequence $(s_k)_k$ converges to s^* a.s. The function h is C^1 in some neighborhood of s^* with first derivatives Lipschitz and J the Jacobian matrix of the mean field h in s^* has all its eigenvalues with negative real part.
- (N2) Let $g_{\hat{\theta}(s)}$ be a solution of the Poisson equation $g - \Pi_{\hat{\theta}(s)} g = H_s - p_{\hat{\theta}(s)}(H_s)$ for any

$s \in \mathcal{S}$. There exists a bounded function w such that

$$w - \Pi_{\hat{\theta}(s^*)} w = g_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)}^T - \Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)} (\Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)})^T - U \quad (20)$$

where the deterministic matrix U is given by :

$$U = \mathbb{E}_{\hat{\theta}(s^*)} \left[g_{\hat{\theta}(s^*)}(z) g_{\hat{\theta}(s^*)}(z)^T - \Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)}(z) \Pi_{\hat{\theta}(s^*)} g_{\hat{\theta}(s^*)}(z)^T \right]. \quad (21)$$

(N3) The step size sequence (γ_k) is decreasing and satisfies $\gamma_k = 1/k^\alpha$ with $2/3 < \alpha < 1$.

Theorem 2 *Under the assumptions of Theorem 1 and under (N1)-(N3), the sequence $(s_k - s^*)/\sqrt{\gamma_k}$ converges in distribution to a Gaussian random vector with zero mean and covariance matrix Γ where Γ is the solution of the following Lyapunov equation:*

$$U + J\Gamma + \Gamma J^T = 0.$$

Moreover,

$$\frac{1}{\sqrt{\gamma_k}}(\theta_k - \theta^*) \rightarrow_{\mathcal{L}} \mathcal{N}(0, \partial_s \hat{\theta}(s^*) \Gamma \partial_s \hat{\theta}(s^*)^T)$$

where $\theta^* = \hat{\theta}(s^*)$.

The proof of Theorem 2 is given in Appendix 7.3.

5 Applications on Bayesian Mixed Effect Template model

5.1 Comparison between MALA and AMALA samplers

As a first experiment, we compare the mixing properties of MALA and AMALA samplers. We used both algorithms to sample from a 10 dimensional normal distribution with zero mean and non diagonal covariance matrix. Its eigenvalues range from 1 to 10. The eigen-directions are chosen randomly. The autocorrelations of both chains are plotted in Fig. 1 where we can see that there is a benefit of using the anisotropic sampler. To evaluate the weight of the anisotropic term $D(x)D(x)^T$ in the covariance matrix, we

compute its amplitude (as its non zero eigenvalue since it is a rank one matrix). We see that it is of the same order as the diagonal part in average and jumps up to 15 times bigger. This shows the importance of the anisotropic term. The last check is the Mean Square Euclidean Jump Distance (MSEJD) which computes the expected squared distance between successive draws of the Markov chain. The two methods provide MSEJD of the same order showing a very slight advantage in term of visiting the space for the AMALA sampler (1.29 versus 1.25 for the MALA).

We will observe in the following experiments that the advantage of considering the AMALA instead of the MALA sampler will be intensified when increasing the problem dimension and including it into our estimation process.

5.2 BME Template estimation

Back to our targeted application, we apply the proposed estimation process on different data bases. The first one is the USPS hand-written digit base as used in [2] and [5]. The other two are medical images of 2D corpus callosum and 3D murine dendrite spine excrescences used in [4].

We begin with presenting the experiments on the USPS database. In order to make comparison, we estimate the parameters in the same conditions as in the previous mentioned works that is to say using the same 20 images per digit. Each image has grey level between 0 (background) and 2 (bright white). These images are presented on the top panel of Fig. 2. We also use a noisy training dataset generated by adding a standardized independent Gaussian noise. These images are presented on the bottom panel of Fig. 2. We test five algorithms: the deterministic approximation of the EM algorithm (FAM-EM) presented in [2], four MCMC-SAEM where the sampler is either the MALA, the adaptive MALA proposed in [7], the hybrid Gibbs sampler presented in [5] and our AMALA algorithm.

For these experiments the tuning parameters are chosen as follows: the threshold b is set to 1,000, the scale δ to 10^{-3} and the regularization ε to 10^{-4} . The other tuning parameters and hyper-parameters are chosen as in [5].

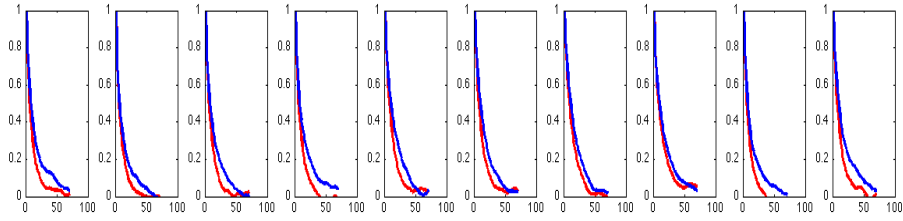


Fig. 1 Autocorrelations of the MALA (blue) and AMALA (red) samplers to target the 10 dimensional normal distribution with anisotropic covariance matrix.

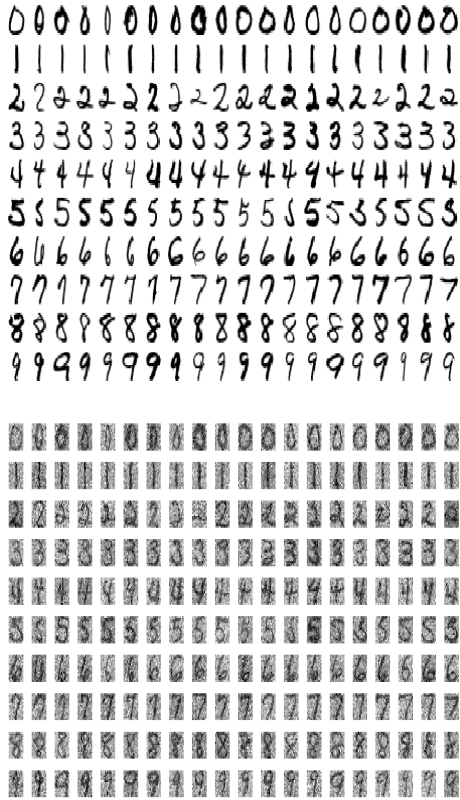


Fig. 2 Top: twenty images per digit of the training set used for the estimation of the model parameters (inverse video). Bottom: same images with additive noise of variance 1.

Note that this model satisfies the conditions of our convergence theorem as these conditions are similar to the ones proved in [5].

5.3 Computational performances

We compare first the computational performances of the algorithms. The computational time is smaller for the three MCMC-SAEM algorithms using "MALA-like" samplers compared to the FAM. Indeed, a numerical convergence of that algorithm requires about 30 to 50 EM steps. Each of them requires a gradient descent which

has 15 iterations in average. This implies to compute 15 times the gradient of the energy (which actually equals our gradient) for each image for each EM step. The "MALA-like"-SAEM algorithms require about 100 to 150 EM steps (depending on the digit) but only one gradient is computed for each image at each step. This reduces the computational time by a factor of at least 4 (up to 7 depending on the digit). No comparison can be done when the data are noisy since the FAM-EM does not converges toward the MAP estimator as mentioned above. Comparing to the hybrid Gibbs-SAEM, the computational time is 8 times lower with the AMALA-SAEM in this particular case of application. Indeed, the hybrid Gibbs sampler requires no computation of the gradient. However, it includes a loop over the coordinates of the hidden variable, here the deformation vector of size $2k_g = 72$. At each of these iterations, the candidate is straightforward to sample whereas the computational cost lies into the acceptance rate. When this becomes heavy, the less times you calculate it, the better. In the AMALA-SAEM, this acceptance rate only has to be calculated once for each image. Therefore, even when the dimension of the hidden variable increases, this is of constant cost. The main price to pay is the computation of the gradient. Therefore, a tradeoff has to be found between the computation of either one gradient or dk_g acceptance rates in order to select the algorithm to use in a given case.

5.4 Results on the template estimation

All the estimated templates obtained with the five algorithms and noise-free and noisy training data are presented in Fig. 3. As noticed in [5], the FAM-EM estimation is sharp when the training set is noise-free and is deteriorated while adding noise. This behavior is not surprising with regard to the theoretical bound

established in [8] in the particular case of compact deformation group. Considering the adaptive sampler, it does not reach a good estimation of the templates which are still very blurry and noisy in both cases. The problem seems to come from the very low acceptance rate already at the beginning of the estimation. The bad initial guess we have about the covariance matrix of the proposal seems to block the chain. Moreover, the tuning parameters are difficult to calibrate along the iterations of the estimation algorithm. Concerning the estimated templates using the Gibbs, MALA and AMALA samplers, they look very similar to each other using the noise-free data as well as the noisy ones. This similarity confirms the convergence of all these algorithms toward the MAP estimator. In this case, the templates are as expected: noise free and sharp.

Nevertheless, when the dimension of the hidden variable increases, both the Gibbs and the MALA samplers show limitations. We run the estimation on the same noisy USPS database, increasing the number k_g of geometrical control points. We choose the dimension of the deformation vector equal to 72, 128 and 200. The Gibbs-SAEM would produce sharp estimations but explodes the computational time. For this reason, we did not run this algorithm on higher dimension experiments. The results are presented in Fig. 4. Concerning the MALA sampler, it does not seem to capture the whole variability of the population in such high dimension. This yields a poorly estimation of the templates. This phenomenon does not appear using our AMALA-SAEM algorithm. The templates still look sharp and the acceptance rates remain reasonable.

5.5 Results on the covariance matrix estimation

Since we are provided with a generative model, once the parameters have been estimated, we can generate synthetic samples in order to evaluate the constrained on the deformations that have been learnt. Some of these samples are presented in Fig. 5. For each digit, 20 examples are generated with the deformations given by $+z$ and 20 others with $-z$ where z is simulated with respect to $\mathcal{N}(0, \Gamma_g)$. We recall that, as already noticed in [5], the Gaussian distribution is symmetric which may lead to strange

samples in one direction whereas the other one looks like something present in the training set.

With regards to the above remarks concerning the computational time and the template estimations, we present in this subsection only the results obtained using MALA and AMALA-SAEM algorithms. We notice that the samples generated by both algorithms look alike in the case of hidden variable of dimension 72. Thus, we present only the results of our AMALA-SAEM estimation. As we can see, the deformations are very well estimated in both cases (without or with noise) and even look similar. This tends to demonstrate that the noise has been separated from the template as well as the geometric variability during the estimation process.

Increasing the dimension of the deformation to 128, we run both algorithms on the noisy dataset. We observe on Fig. 6 that the geometric variability of the samples remains similar to the one obtained in lower dimension using our AMALA-SAEM. However, the MALA-SAEM does not manage to capture the whole variability of the deformations which is related to the results observed above on the template. This confirms the limitation of the use of MALA-SAEM in higher dimension.

5.6 Results on the noise variance estimation

The last check of the accuracy of the estimation relies in the noise variance estimation. The plots of their evolutions along the AMALA-SAEM iterations for each digit in both cases (without and with noise) are presented in Fig. 7. This variance is underestimated in particular in the noisy case, which is a well-known effect of the maximum likelihood estimator. We observe that the geometrically very constrained digits as 1 or 7 tend to converge very quickly whereas the digits 2 and 4 require more iterations to capture all the shape variability.

Since this is a real parameter, we used it to illustrate the Central Limit Theorem stated in Subsection 4.3. Figure 8 and Figure 9 show the histograms of 10,000 runs of the algorithm with the same initial conditions. We use the digits 0 and 2 of the original data set as well as of the noisy data. As the iterations go along, the distribution of the estimates tends to look like a Gaussian distribution centered in the estimated

Algo./ Noise	FAM	Hybrid Gibbs	MALA	Adaptive MALA	AMALA
No Noise					
Noise					

Fig. 3 Estimated templates using the five algorithms on noise free and noisy data. The training set includes 20 images per digit. The dimension of the hidden variable is 72.

Dim. of def. / Sampler	$2k_g = 72$	$2k_g = 128$	$2k_g = 200$
MALA			
AMALA			

Fig. 4 Estimated templates using MALA and AMALA samplers in the stochastic EM algorithm on noisy training data. The training set includes 20 images per digit. The dimension of the hidden variable increases from 72 to 200.

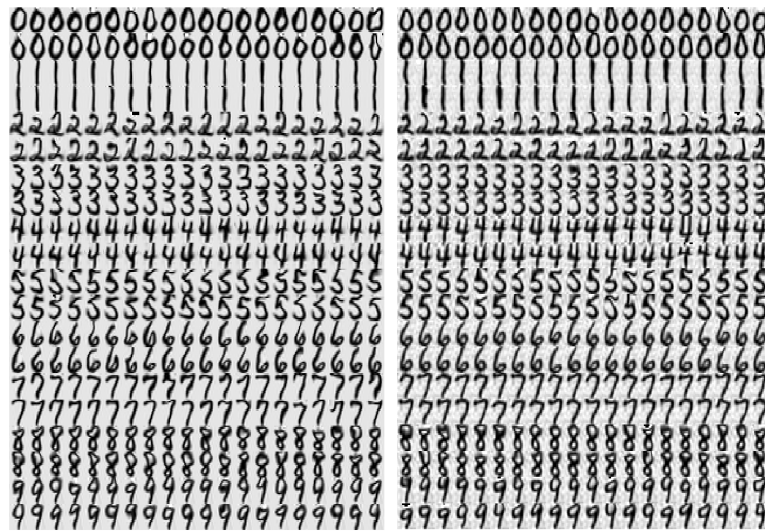


Fig. 5 Synthetic samples generated with respect to the BME template model using the estimated parameters with AMALA-SAEM. For each digit, the two lines represent the deformation using + and - the simulated deformation z . Left: data without noise. Right: data with noise variance 1. The number of geometric control points is 36 leading to a hidden variable of dimension 72.

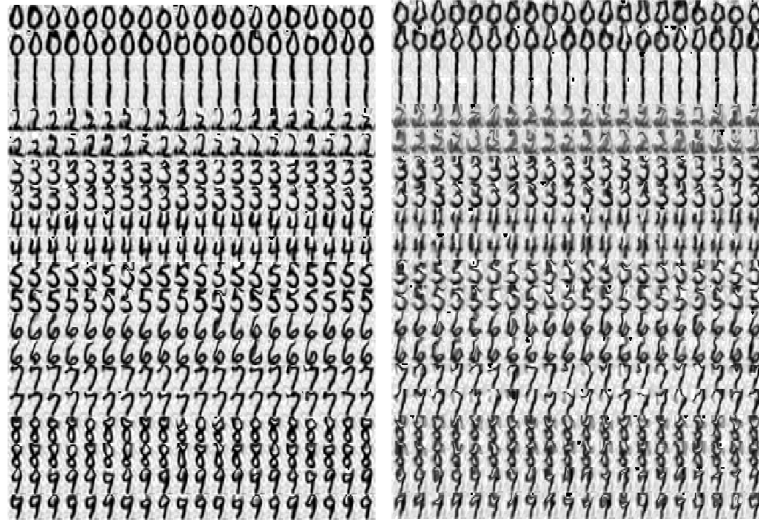


Fig. 6 Synthetic samples generated with respect to the BME template model using the estimated parameters with AMALA-SAEM (left) and MALA-SAEM (right). For each digit, the two lines represent the deformation using $+$ and $-$ the simulated deformation z . The number of geometric control points is 64 leading to a hidden variable of dimension 128.

noise variances which demonstrates empirically the Central Limit Theorem.

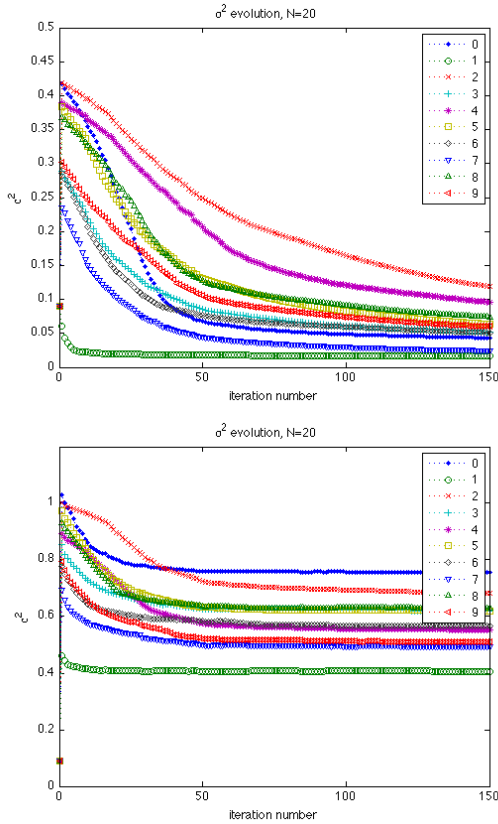


Fig. 7 Evolution of the estimation of the noise variance along the AMALA-SAEM iterations. Top: original data. Bottom: noisy data.

5.7 Classification results

The deformable template model enables to perform classification using the maximum likelihood of a new image to allocate it to one class, here the digit. We use the test USPS database (which contains 2007 digits) for classification while the training was done on the previous 20 noisy images. The results obtained with the hybrid Gibbs, MALA and AMALA-SAEM are presented in Table 1. In dimension 72, the best classification rate is performed by the hybrid Gibbs-SAEM. This is easily understandable since the sampling scheme enables to catch deformations which have been optimized control point by control point. Therefore, the estimated covariance matrix carries more local accuracy. The AMALA-SAEM leading to a much smaller computation time and to estimates of the same quality provides also a very good classification rate. This confirms the good results observed on both the template estimates and the synthetic samples. Unfortunately, the MALA-SAEM shows again some limitations. Even if the templates look acceptable, the sampler does not manage to capture the whole class variability. Therefore, the classification rate falls down.

In order to evaluate the stability of our estimation algorithm with respect to the dimension, we perform the same classification with more control points. As expected, the MALA-SAEM classification rate is deteriorated whereas our AMALA-SAEM keeps very good perfor-

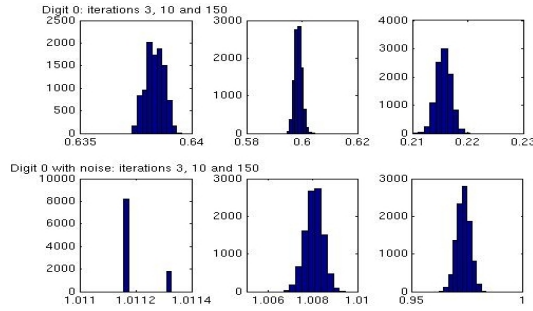


Fig. 8 Empirical convergence toward the Gaussian distribution of the estimated noise variance along the AMALA-SAEM iterations for digit 0. Top: original data. Bottom: noisy data.

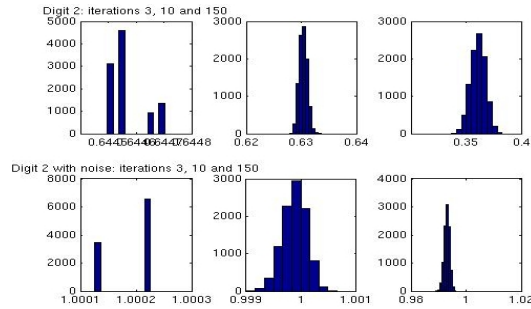


Fig. 9 Empirical convergence toward the Gaussian distribution of the estimated noise variance along the AMALA-SAEM iterations for digit 2. Top: original data. Bottom: noisy data.

mances. Note that the hybrid Gibbs sampler was not tested in dimension $2k_g = 128$ because of its very long computational time.

Sampler / Dim. of def.	Hybrid Gibbs	MALA	AMALA
72	22.43	35.98	23.22
128	×	43.8	25.36

Table 1 Error rate using the estimations on the noisy training set with respect to the sampler used in the MCMC-SAEM algorithm and the dimension of the deformation $2k_g$. The classification is performed on the test set of the USPS database.

5.8 2D medical image template estimation

A second database is used to illustrate our algorithm. As before, in order to make comparisons with existing algorithms, we use the same database presented in [4]. It consists of 47 medical images, each of them is a $2D$ square zone around the end point of the corpus callosum. This box contains a part of this corpus callosum as well as a part of the cerebellum. Ten exemplars are presented in the top rows of Fig. 10.

The estimations are compared with these obtained with the FAM-EM and the hybrid Gibbs-SAEM algorithms and with the grey level mean image (bottom row of Fig. 10). In this real situation, the Euclidean grey level mean image (a) is very blurry. The estimated template using the FAM-EM (b) provides a first amelioration in particular leading to a sharper corpus callosum. However, the cerebellum still looks blurry in particular when comparing it to the shape which appears in the template estimated using the hybrid Gibbs SAEM (c). The result of our AMALA-SAEM is given in image (d). This template is very close to (c) as we could expect at a convergence point. Nevertheless the AMALA-SAEM has much lower computational time than the hybrid Gibbs-SAEM. This shows the advantage of using AMALA-SAEM in real cases of high dimension.

5.9 3D medical image template estimation

We also test our algorithm in much higher dimension using the dataset of murine dendrite spines (see [1,9,10]) already used in [4]. The dataset consists of 50 binary images of microscopic structures, tiny protuberances found on many types of neurons termed dendrite spines.

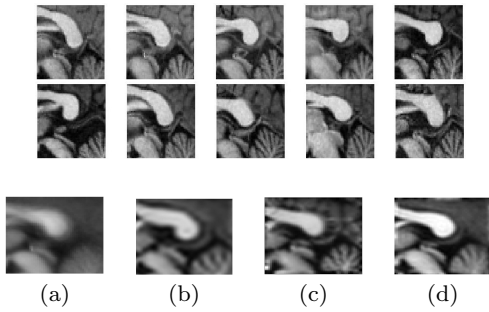


Fig. 10 Medical image template estimation. Top rows : 10 Corpus callosum and cerebellum training images among the 47 available. Bottom row : (a) mean image. (b) FAM-EM estimated template. (c) Hybrid Gibbs - SAEM estimated template. (d) AMALA-SAEM estimated template.

The images are from control mice and knock-out mice which have been genetically modified to mimic human neurological pathologies like Parkinson’s disease. The acquisition process consisted of electron microscopy after injection of Lucifer yellow and subsequent photo-oxidation. The shapes were then manually segmented on the tomographic reconstruction of the neurons. Some of these binary images are presented in Fig. 11 which shows a 3D view of some exemplars among the training set. Each image is a binary (background = 0, object = 2) cubic volume of size 28^3 . We can notice here the large geometrical variability of this population of images. Therefore we use a hidden variable of dimension $3k_g = 648$ to catch this complex structure.

The template estimated with either 30 or 50 observations are presented in Fig. 13. We obtain similar shapes which are coherent with what a mean shape could be regarding the training sample. To evaluate the estimated geometrical variability, we generate synthetic samples as done in Subsection 5.5. Eight of these are shown in Fig. 12. We observe different twisting which are all coherent with the shapes observed in the dataset. Note that the training shapes have very irregular boundaries whereas the parametric model used for the template leads to a smoother image. Thus, the synthetic samples do not reflect the local ruggedness of the segmented murine dendrite spines. If the aim was to capture these local bumps, the number of photometrical control points has to be increased. However, the goal of our study was to detect global shape deformations.



Fig. 13 Estimated templates of murine dendrite spines. The training set is either composed of 30 (left) or 50 (right) images.

6 Conclusion

In this paper we have considered the deformable template estimation issue using the BME model. We were particularly interested in the high dimensional setting. To that purpose, we have proposed to optimize the sampling scheme in the MCMC-SAEM algorithm to get an efficient and accurate estimation process. We have exhibited a new MCMC method based on the classical Metropolis Adjusted Langevin Algorithm where we introduced an anisotropic covariance matrix in the proposal. This optimization takes into account the anisotropy of the target distribution. We proved that the generated Markov chain is geometrically ergodic uniformly on any compact set. We have also proved the almost sure convergence of the sequence of parameters generated by the estimation algorithm as well as its asymptotic normality. We have illustrated this estimation algorithm in the BME model. We have considered different datasets of the literature namely the USPS database, 2D medical images of corpus callosum and 3D medical images of murine dendrite excrescences. We have compared the results with previously published ones to highlight the gain in speed and accuracy of the proposed algorithm.

We emphasize that the proposed estimation scheme can be applied in a wide range of application fields involving missing data models in high dimensional setting. In particular, this method is promising when considering mixture models as proposed in [3]. Indeed, it will enable to shorten the computation time of the simulation part which in that case requires the use of many auxiliary Markov chains. This also provides a good tool for this BME model when introducing a diffeomorphic constrain on the deformations. In this case, it is even more important to get an efficient estimation process since the computational cost of diffeomorphic deformation is intrinsically large.

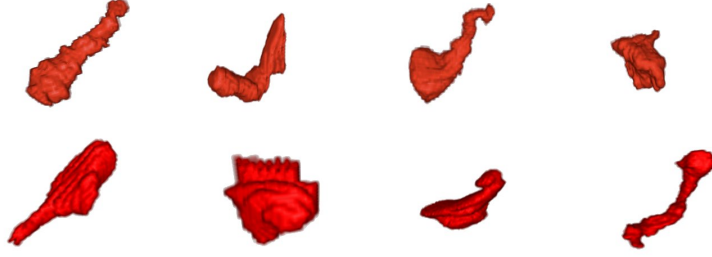


Fig. 11 3D views of eight samples of the data set of dendrite spines. Each image is a volume leading to a binary image.

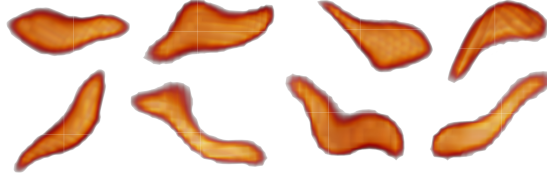


Fig. 12 3D views of eight synthetic data. The estimated template shown on the left of Fig. 13 is randomly deformed with respect to the estimated covariance matrix.

7 Appendix

7.1 Proof of Proposition 1

The idea of the proof is the same as the one of the geometric ergodicity of the random walk Metropolis algorithm developed in [19] and reworked in [7] for its adaptive version of the MALA with truncated drift. The fact that both the drift and the covariance matrix are bounded even depending on the gradient of $\log \pi_s$ enables partially similar proofs.

Let us first recall the transition kernel:

$$\begin{aligned} \Pi_s(x, A) &= \int_A \alpha_s(x, z) q_s(x, z) dz + \\ &\quad \mathbf{1}_A(x) \int_{\mathcal{X}} (1 - \alpha_s(x, z)) q_s(x, z) dz, \end{aligned} \quad (22)$$

where $\alpha_s(x, z) = \min(1, \rho_s(x, z))$ and $\rho_s(x, z) = \frac{\pi_s(z) q_s(z, x)}{q_s(x, z) \pi_s(x)}$.

Thanks to the bounded drift and covariance matrix, we can bound uniformly in $s \in \mathcal{S}$ the proposal distribution q_s by two centered Gaussian distributions as follows: there exist constants $0 < k_1 < k_2$, $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that for all $(x, z) \in \mathcal{X}^2$ and for all $s \in \mathcal{S}$

$$k_1 g_{\epsilon_1}(x - z) \leq q_s(x, z) \leq k_2 g_{\epsilon_2}(x - z), \quad (23)$$

denoting by g_a the centered Gaussian probability density function in \mathbb{R}^l with covariance matrix aId_l .

7.1.1 Proof of the existence of a small set \mathcal{C}

Let \mathcal{C} be a compact subset of \mathcal{X} .

Let K be a compact set. We define $\tau = \inf\{\rho_s(x, z), x \in \mathcal{C}, z \in K, s \in \mathcal{K}\}$. Since ρ_s is a ratio of positive continuous functions in s, x and z and \mathcal{K} is a compact subset of \mathcal{S} , we have $\tau > 0$. The same argument holds for $(s, x, z) \mapsto q_s(x, z)$ which is bounded by below by $\mu > 0$. Therefore, for all $x \in \mathcal{C}$, for any $A \in \mathcal{B}$ and for all $s \in \mathcal{K}$:

$$\begin{aligned} \Pi_s(x, A) &\geq \int_{A \cap K} \alpha_s(x, z) q_s(x, z) dz \\ &\geq \min(1, \tau) \mu \int_A \mathbf{1}_K(z) dz. \end{aligned}$$

Therefore, we can define $\nu(A) = \frac{1}{Z} \int_A \mathbf{1}_K(z) dz$ where Z is the renormalisation constant and $\epsilon = \min(1, \tau) \mu Z$ so that \mathcal{C} is a small set for the transition kernel Π_s for all $s \in \mathcal{K}$ and (17) holds.

7.1.2 Proof of the drift condition

We will prove this property in two steps. First, we establish that each kernel Π_s satisfies a Drift property with a specific function V_s . Then, we construct a common function V so that we will be able to prove the Drift property uniformly in $s \in \mathcal{K}$.

Let us concentrate on the first step. Let us consider s fixed. As already suggested in [19],

we only need to prove the two following conditions:

$$\sup_{x \in \mathcal{X}} \frac{\Pi_s V_s(x)}{V_s(x)} < \infty \quad (24)$$

and

$$\limsup_{|x| \rightarrow \infty} \frac{\Pi_s V_s(x)}{V_s(x)} < 1. \quad (25)$$

We take the same path as in [7] applied to our case and refer to Fig. 14 for a 2D visualization of all the sets introduced along the proof.

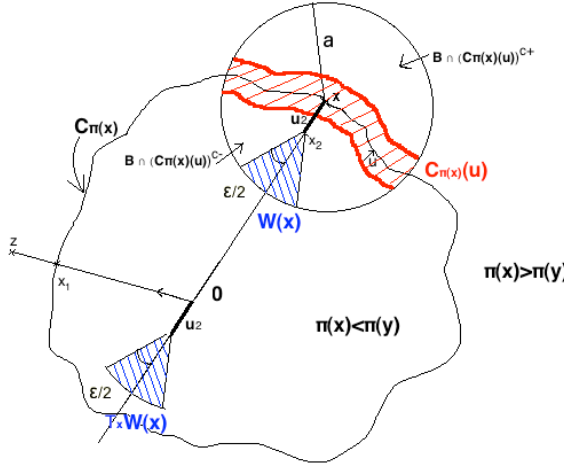


Fig. 14 2D representation of the sets used in the proof.

For any $x \in \mathcal{X}$, we denote by $A_s(x) = \{z \in \mathcal{X} \text{ such that } \rho_s(x, z) \geq 1\}$ the acceptance set and by $R_s(x) = A_s(x)^c$ its complementary set. Then, we recall $V_s(x) = c_s \pi_s(x)^{-\beta}$ for some $\beta \in]0, 1[$. Therefore, for all $x \in \mathcal{X}$:

$$\begin{aligned} \frac{\Pi_s V_s(x)}{V_s(x)} &= \int_{A_s(x)} q_s(x, z) \frac{V_s(z)}{V_s(x)} dz + \\ &\int_{R_s(x)} \frac{\pi_s(z) q_s(z, x)}{\pi_s(x) q_s(x, z)} q_s(x, z) \frac{V_s(z)}{V_s(x)} dz + \\ &\int_{R_s(x)} \left(1 - \frac{\pi_s(z) q_s(z, x)}{\pi_s(x) q_s(x, z)}\right) q_s(x, z) dz \\ &\leq \int_{A_s(x)} \underbrace{\frac{\pi_s(z)^{-\beta}}{\pi_s(x)^{-\beta}} q_s(x, z)}_{f_1(x, z)} dz + \\ &\int_{R_s(x)} \underbrace{\frac{\pi_s(z)^{1-\beta}}{\pi_s(x)^{1-\beta}} q_s(z, x)}_{f_2(x, z)} dz + \int_{R_s(x)} \underbrace{q_s(x, z)}_{f_3(x, z)} dz. \end{aligned}$$

On the acceptance set $A_s(x)$, we have:

$$\frac{\pi_s(z)^{-\beta}}{\pi_s(x)^{-\beta}} q_s(x, z) \leq q_s(z, x)^\beta q_s(x, z)^{1-\beta}.$$

Thanks to Equation (23) one can bound this right hand side by the following symmetric Gaussian distribution:

$$\frac{\pi_s(z)^{-\beta}}{\pi_s(x)^{-\beta}} q_s(x, z) \leq k_2 g_{\epsilon_2}(z - x) \quad (26)$$

which yields:

$$\int_{A_s(x)} f_1(x, y) dz \leq k_2 \int_{A_s(x)} g_{\epsilon_2}(z - x) dz. \quad (27)$$

Equivalently on $R_s(x)$, we have the following bound:

$$\begin{aligned} \frac{\pi_s(z)^{1-\beta}}{\pi_s(x)^{1-\beta}} q_s(z, x) &\leq q_s(x, z)^{1-\beta} q_s(z, x)^\beta \quad (28) \\ &\leq k_2 g_{\epsilon_2}(z - x). \quad (29) \end{aligned}$$

Let fix $\varepsilon > 0$, there exists $a > 0$ such that $\int_{B(x, a)} g_{\epsilon_2}(z - x) dz \geq 1 - \varepsilon$. This leads to:

$$\int_{A_s(x) \cap B(x, a)^c} f_1(x, z) dz \leq k_2 \varepsilon.$$

Let $C_{\pi_s(x)}$ be the level set of π_s in x : $C_{\pi_s(x)} = \{z \in \mathcal{X} : \pi_s(z) = \pi_s(x)\}$. We define a pipe around this level set as

$$C_{\pi_s(x)}(u) = \{z + t n(z), |t| \leq u, z \in C_{\pi_s(x)}\}.$$

Thanks to assumption **(B1)**, there exists $r_1 > 0$ such that for all $x \in \mathcal{X}$ satisfying $|x| \geq r_1$ then 0 is inside the hyperspace defined by the level set $C_{\pi_s(x)}$ ($\pi_s(0) > \pi_s(x)$). Therefore, let $x \in \mathcal{X}$, $|x| \geq r_1$, then for all $z \in \mathcal{X}$, $\exists x_1 \in C_{\pi_s(x)}$ and $t > 0$ such that $z = x_1 + t n(x_1)$.

Since $z \mapsto g_{\epsilon_2}(z - x)$ is a smooth density in the variable z , we can find $u > 0$ sufficiently small such that

$$\int_{B(x, a) \cap C_{\pi_s(x)}(u)} g_{\epsilon_2}(z - x) dz \leq \varepsilon, \quad (30)$$

leading to

$$\int_{A_s(x) \cap B(x, a) \cap C_{\pi_s(x)}(u)} f_1(x, z) dz \leq k_2 \varepsilon.$$

Assumption **(B1)** implies that for any $r > 0$ and $t > 0$, $d_r(t) = \sup_{|x| \geq r} \frac{\pi_s(x+t n(x))}{\pi_s(x)}$ goes to 0 as r goes to ∞ . Denote $C_{\pi_s(x)}(u)^{c+} = \{z \in C_{\pi_s(x)}(u)^c \text{ s.t. } \pi_s(x) > \pi_s(z)\}$ and $C_{\pi_s(x)}(u)^{c-} = \{z \in C_{\pi_s(x)}(u)^c \text{ s.t. } \pi_s(x) < \pi_s(z)\}$. We denote

$\mathcal{D}^+ = A_s(x) \cap B(x, a) \cap C_{\pi_s(x)}(u)^{c+}$. Therefore there exists $r_2 > r_1 + a$ such that for any x , $|x| \geq r_2$

$$\begin{aligned} \int_{\mathcal{D}^+} f_1(x, z) dz &\leq \int_{\mathcal{D}^+} \left(\frac{\pi_s(z)}{\pi_s(x)} \right)^{1-\beta} q_s(z, x) dz \\ &\leq d_{r_2}(u)^{1-\beta} k_2 \int_{\mathcal{X}} g_{\varepsilon_2}(z - x) dz \\ &\leq k_2 d_{r_2}(u)^{1-\beta}, \end{aligned}$$

using Equation (15) which states that the stationary distribution is decreasing in the direction of the normal of x sufficiently large.

In the same way, one has on the set $\mathcal{D}^- = A_s(x) \cap B(x, a) \cap C_{\pi_s(x)}(u)^{c-}$

$$\begin{aligned} \int_{\mathcal{D}^-} f_1(x, z) dz &\leq \int_{\mathcal{D}^-} \left(\frac{\pi_s(z)}{\pi_s(x)} \right)^{-\beta} q_s(x, z) dz \\ &\leq k_2 d_{r_2}(u)^\beta. \end{aligned}$$

The same inequalities can be obtained for f_2 using the same arguments:

$$\begin{aligned} \int_{R_s(x) \cap B(x, a)^c} f_2(x, z) dz &\leq k_2 \varepsilon \\ \int_{R_s(x) \cap B(x, a) \cap C_{\pi_s(x)}(u)} f_2(x, z) dz &\leq k_2 \varepsilon \\ \int_{R_s(x) \cap B(x, a) \cap C_{\pi_s(x)}(u)^{c+}} f_2(x, z) dz &\leq k_2 d_{r_2}(u)^{1-\beta} \\ \int_{R_s(x) \cap B(x, a) \cap C_{\pi_s(x)}(u)^{c-}} f_2(x, z) dz &\leq k_2 d_{r_2}(u)^\beta. \end{aligned}$$

This yields

$$\limsup_{|x| \rightarrow \infty} \frac{\Pi_s V_s(x)}{V_s(x)} \leq \limsup_{|x| \rightarrow \infty} \int_{R_s(x)} q_s(x, z) dz. \quad (31)$$

Let $Q(x, A_s(x)) = \int_{A_s(x)} q_s(x, z) dz$, we get

$$\limsup_{|x| \rightarrow \infty} \frac{\Pi_s V_s(x)}{V_s(x)} \leq 1 - \liminf_{|x| \rightarrow \infty} Q(x, A_s(x)).$$

Let us now prove that

$\liminf_{|x| \rightarrow \infty} Q(x, A_s(x)) \geq c > 0$ where c does not depend on x .

Let a fixed as above. Since q_s is an exponential function, there exists $c_0^a > 0$ such that for all $x \in \mathcal{X}$ and $s \in \mathcal{S}$,

$$\inf_{z \in B(x, a)} \frac{q_s(z, x)}{q_s(x, z)} \geq c_0^a. \quad (32)$$

Moreover, thanks to assumption **(B1)** there exists $r_3 > 0$ such that for all $x \in \mathcal{X}$, $|x| \geq r_3$, there exists $0 < u_2 < a$ such that,

$$\frac{\pi_s(x)}{\pi_s(x - u_2 n(x))} \leq c_0^a. \quad (33)$$

Hence, for $|x| \geq r_3$, any point $x_2 = x - u_2 n(x)$ belongs to $A_s(x)$.

Let $W(x)$ be the cone defined as:

$$W(x) = \left\{ x_2 - t\zeta, 0 < t < a - u_2, \zeta \in \mathcal{S}^{d-1}, |\zeta - n(x_2)| \leq \frac{\varepsilon}{2} \right\} \quad (34)$$

where \mathcal{S}^{d-1} is the unit sphere in \mathbb{R}^d .

Let us prove that $W(x) \subset A_s(x)$.

Using assumption **(B1)**, we have for a sufficiently large x : $m(x) \cdot n(x) \leq -\varepsilon$. Besides, by construction of $W(x)$ for large x , for all $z \in W(x)$, $|n(z) - n(x)| \leq \varepsilon/2$ with $n(x) = n(x_2)$ (see Fig. 14). This leads to for any sufficiently large x , for all $z \in W(x)$,

$$\begin{aligned} m(z) \cdot \zeta &= m(z) \cdot (\zeta - n(x_2)) + m(z) \cdot (n(x_2) - n(z)) \\ &+ m(z) \cdot n(z) \leq \varepsilon/2 + \varepsilon/2 - \varepsilon = 0. \end{aligned} \quad (35)$$

Let now $z = x_2 - t\zeta \in W(x)$. Using the mean value theorem on the differentiable function π_s between x_2 and z , we get that there exists $\tau \in]0, s[$ such that $\pi_s(z) - \pi_s(x_2) = -t\zeta \cdot \nabla \pi_s(x_2 - \tau\zeta)$. Using the definition of m , this implies that $\pi_s(z) - \pi_s(x_2) = -t\zeta \cdot m(x_2 - \tau\zeta) |\nabla \pi_s(x_2 - \tau\zeta)| \geq 0$ thanks to Equation (35). Putting all these results together we finally get that for all $z \in W(x)$, $\pi_s(z) \geq \pi_s(x_2) \geq \frac{1}{c_0^a} \pi_s(x)$. Moreover, as $z \in B(x, a)$ as well, Equation (32) is satisfied, leading to $z \in A_s(x)$.

Then, we have

$$\begin{aligned} Q(x, A_s(x)) &= \int_{A_s(x)} q_s(x, z) dz \\ &\geq \int_{A_s(x)} k_1 g_{\varepsilon_1}(z - x) dz \\ &\geq k_1 \int_{W(x)} g_{\varepsilon_1}(z - x) dz \\ &= \int_{T_x(W(x))} g_{\varepsilon_1}(z) dz \end{aligned}$$

where

$$T_x(W(x)) = \left\{ -u_2 n(x) - t\zeta, 0 < t < a - u_2, \zeta \in \mathcal{S}^{d-1}, |\zeta - n(x)| \leq \frac{\varepsilon}{2} \right\} \quad (36)$$

is the translation of the set $W(x)$ by the vector x . Note that $W(x)$ does not depend on s . But since g_{ε_1} is isotropic and $T_x(W(x))$ only depends on a fixed constant u_2 and $n(x)$, this last integral is independent of x , so there exists

a positive constant c independent of $s \in \mathcal{S}$ such that:

$$c = \int_{T_x(W(x))} g_{\epsilon_1}(z) dz. \quad (37)$$

Back to our limit, for all $s \in \mathcal{S}$

$$\limsup_{|x| \rightarrow \infty} \frac{\Pi_s V_s(x)}{V_s(x)} \leq 1 - c \quad (38)$$

which ends the proof of the condition (25).

To prove (24), we use the previous result. Indeed, since $\frac{\Pi_s V_s(x)}{V_s(x)}$ is a smooth function on \mathcal{X} it is bounded on every compact subset. Moreover since the limsup is finite, then it is also bounded outside a fixed compact. This proves the results.

Thanks to assumption **(B2)** and the bounded drift for all $s \in \mathcal{S}$, there exists a constant c_0^a uniform in $s \in \mathcal{S}$ such that Equations (32) and (33) still hold for all $s \in \mathcal{S}$. This implies, as mentioned above, that the set $T_x(W(x))$ is independent of $s \in \mathcal{S}$. Therefore, we can set $\tilde{\lambda} = 1 - c < 1$ where c is defined in Equation (37) and is also independent of $s \in \mathcal{S}$.

This proves the Drift property for the function V_s : there exist constants $0 < \tilde{\lambda} < 1$ and $\tilde{b} > 0$ such that for all $x \in \mathcal{X}$,

$$\Pi_s V_s(x) \leq \tilde{\lambda} V_s(x) + \tilde{b} \mathbf{1}_{\mathcal{C}}(x), \quad (39)$$

where \mathcal{C} is a small set. Note that \tilde{b} is also independent of $s \in \mathcal{S}$ using the same arguments as before.

Let us now exhibit a function V independent of $s \in \mathcal{S}$ and prove the uniform Drift condition.

We define for all $x \in \mathcal{X}$,

$$V(x) = V_1(x)^\xi V_2(x)^{2\xi} \quad (40)$$

for $0 < \xi < \min(1/2\beta, b_0/4)$. Therefore, for all $s \in \mathcal{S}$, for all $\varepsilon > 0$ we have,

$$\begin{aligned} \Pi_s V(x) &= \int_{\mathcal{X}} \Pi_s(x, z) V_1(z)^\xi V_2(z)^{2\xi} dz \\ &\leq \frac{1}{2} \int_{\mathcal{X}} \Pi_s(x, z) \left(\frac{V_1(z)^{2\xi}}{\varepsilon^2} + \varepsilon^2 V_2(z)^{4\xi} \right) dz \\ &\leq \frac{1}{2\varepsilon^2} \int_{\mathcal{X}} \Pi_s(x, z) V_s(z)^{2\xi} dz + \\ &\quad \frac{\varepsilon^2}{2} \int_{\mathcal{X}} \Pi_s(x, z) V_2(z)^{4\xi} dz. \quad (41) \end{aligned}$$

Applying the Drift property for Π_s with $V_s^{2\xi}$,

$$\begin{aligned} \Pi_s V(x) &\leq \frac{1}{2\varepsilon^2} (\tilde{\lambda} V_s(x)^{2\xi} + \tilde{b} \mathbf{1}_{\mathcal{C}}(x)) \\ &\quad + \frac{\varepsilon^2}{2} \int_{\mathcal{X}} \Pi_s(x, z) V_2(z)^{4\xi} dz. \quad (42) \end{aligned}$$

Using the definition of V and the fact that V_1 is bounded by below by 1, we get:

$$\begin{aligned} \Pi_s V(x) &\leq \frac{\tilde{\lambda}}{2\varepsilon^2} V(x) + \frac{\tilde{b}}{2\varepsilon^2} \mathbf{1}_{\mathcal{C}}(x) \\ &\quad + \frac{\varepsilon^2}{2} \int_{\mathcal{X}} \Pi_s(x, z) V_2(z)^{4\xi} dz. \quad (43) \end{aligned}$$

Since $0 < \tilde{\lambda} < 1$ is independent of $s \in \mathcal{S}$ and using assumption **(B3)**, there exists $\xi > 0$ such that

$$\sup_{s \in \mathcal{S}, x \in \mathcal{X}} \int_{\mathcal{X}} \Pi_s(x, z) V_2(z)^{4\xi} dz \leq \frac{2}{1 + \tilde{\lambda}}. \quad (44)$$

This yields

$$\Pi_s V(x) \leq \left(\frac{\tilde{\lambda}}{2\varepsilon^2} + \frac{\varepsilon^2}{1 + \tilde{\lambda}} \right) V(x) + \frac{\tilde{b}}{2\varepsilon^2} \mathbf{1}_{\mathcal{C}}(x). \quad (45)$$

We can now fix $\varepsilon^2 = \sqrt{\frac{\tilde{\lambda}(1+\tilde{\lambda})}{2}}$ which leads to

$$\Pi_s V(x) \leq \sqrt{\frac{2\tilde{\lambda}}{1+\tilde{\lambda}}} V(x) + \frac{\tilde{b}}{2\varepsilon^2} \mathbf{1}_{\mathcal{C}}(x). \quad (46)$$

We set $\lambda = \sqrt{\frac{2\tilde{\lambda}}{1+\tilde{\lambda}}} < 1$ and $b = \frac{\tilde{b}}{2\varepsilon^2} > 0$ which concludes the proof.

7.2 Proof of Theorem 1

We provide here the proof of the convergence of the estimated sequence generated by Algorithm 1.

We apply Theorem 4.1 from [5] with the functions H_s equals to $H_s(z) = S(z) - s$, $\Pi_s = \Pi_{\hat{\theta}(s)}$, $\pi_s = p_{\hat{\theta}(s)}$ and

$$h(s) = \int (S(z) - s) p_{\hat{\theta}(s)}(z) \mu(dz).$$

Let us first prove assumption **(A1')** which ensures the existence of a global Lyapunov function for the mean field of the stochastic approximation. It guaranties that, under some conditions, the sequence $(s_k)_{k \geq 0}$ remains in a compact subset of \mathcal{S} and converges to the set of critical points of the log-likelihood.

Assumptions **(M1)**-**(M7)** ensure that \mathcal{S} is an open subset and that the function h is continuous on \mathcal{S} . Moreover defining $w(s) = -l(\hat{\theta}(s))$, we get that w is continuously differentiable on \mathcal{S} . Applying Lemma 2 of [13], we get **(A1')(i)**, **(A1')(iii)** and **(A1')(iv)**.

To prove **(A1')(ii)**, we consider as absorbing set \mathcal{S}_a the closure of the convex hull of $S(\mathbb{R}^l)$ denoted $\text{Conv}(S(\mathbb{R}^l))$. So assumption **(M7)(ii)** is exactly equivalent to assumption **(A1')(ii)**.

This achieves the proof of assumption **(A1')**.

Let us now prove assumption **(A2)** which states in particular the existence of a unique invariant distribution for the Markov chain.

To that purpose, we prove that our family of kernels satisfies the drift conditions mentioned in [6] and used in [5] in a similar context. These conditions are the existence of a small set uniformly in $s \in \mathcal{K}$, the uniform drift condition and an upper bound on the family kernel :

(DR1) For any $s \in \mathcal{S}$, $\Pi_{\hat{\theta}(s)}$ is ψ -irreducible and aperiodic. In addition there exist a function $V : \mathbb{R}^l \rightarrow [1, \infty[$ and a constant $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \mathcal{S}$, there exist an integer j and constants $0 < \lambda < 1$, $B, \kappa, \delta > 0$ and a probability measure ν such that

$$\sup_{s \in \mathcal{K}} \Pi_{\hat{\theta}(s)}^j V^p(z) \leq \lambda V^p(z) + B \mathbf{1}_{\mathcal{C}}(z), \quad (47)$$

$$\sup_{s \in \mathcal{K}} \Pi_{\hat{\theta}(s)} V^p(z) \leq \kappa V^p(z) \quad \forall z \in \mathcal{X}, \quad (48)$$

$$\inf_{s \in \mathcal{K}} \Pi_{\hat{\theta}(s)}^j(z, A) \geq \delta \nu(A) \quad \forall z \in \mathcal{C}, \forall A \in \mathcal{B}(\mathcal{C}) \quad (49)$$

Let us start with the irreducibility of $\Pi_{\hat{\theta}(s)}$. The kernel $\Pi_{\hat{\theta}(s)}$ is bounded by below as follows :

$$\Pi_{\hat{\theta}(s)}(x, A) \geq \int_A \alpha_s(x, z) q_s(x, z) dz, \quad (50)$$

where $\alpha_s(x, z) = \min(1, \rho_s(x, z))$ and $\rho_s(x, z) = \frac{\pi_s(z) q_s(z, x)}{q_s(x, z) \pi_s(x)} > 0$. Since the proposal density q_s is positive, this proves that $\Pi_s(x, A)$ is positive and the ψ -irreducibility of each kernel of the family.

Proposition 1 and Remark 1 show that Equations (47) and (49) hold for $j = 1$ with V defined in Equation (40) and some $p > 2$. Moreover, since Equation (47) holds for $j = 1$ and $V \geq 1$, Equation (48) directly comes from Equation (47) choosing $\kappa = B + \lambda$. This implies all three inequalities. Since the small set condition

is satisfied with $j = 1$ (small set "in one-step"), each chain of the family is aperiodic (see [25]).

Assumption **(A2)** is therefore directly implied by assumption **(M1)**.

Let us now prove assumption **(A3')** which states some regularity conditions (Hölder type ones) on the solution of the Poisson equation related to the transition kernel. It also ensures that this solution and its image through the transition kernel have reasonable behaviors as the chain goes to infinity and that the kernel is V^p -bounded in expectation.

The drift conditions proved previously imply the geometric ergodicity uniformly in s in any compact set \mathcal{K} . This also ensures the existence of a solution of the Poisson equation (see [25]) required in Assumption **(A3')**.

We first consider condition **(A3'(i))**.

Let us define for any $g : \mathcal{X} \rightarrow \mathbb{R}^m$ the norm $\|g\|_V \triangleq \sup_{z \in \mathcal{X}} \frac{\|g(z)\|}{V(z)}$.

Since $H_s(z) = S(z) - s$, assumptions **(M8)** and **(B1)** ensure that $\sup_{s \in \mathcal{K}} \|H_s\|_V < \infty$ and inequality (4.3) of **(A3'(i))** holds.

The uniform ergodicity of the family of Markov chains corresponding to the AMALA on \mathcal{K} ensures that there exist constants $0 < \gamma_{\mathcal{K}} < 1$ and $C_{\mathcal{K}} > 0$ such that for all $s \in \mathcal{K}$

$$\begin{aligned} \sup_{s \in \mathcal{K}} \|g_{\hat{\theta}(s)}\|_V &= \sup_{s \in \mathcal{K}} \left\| \sum_{k \geq 0} (\Pi_{\hat{\theta}(s)}^k H_s - p_{\hat{\theta}(s)} H_s) \right\|_V \\ &\leq \sup_{s \in \mathcal{K}} \sum_{k \geq 0} C_{\mathcal{K}} \gamma_{\mathcal{K}}^k \|H_s\|_V < \infty. \end{aligned}$$

Thus for all s in \mathcal{K} , $g_{\hat{\theta}(s)}$ belongs to $\mathcal{L}_V = \{g : \mathbb{R}^l \rightarrow \mathbb{R}^m, \|g\|_V < \infty\}$.

Repeating the same calculation as above, it is immediate that $\sup_{s \in \mathcal{K}} \|\Pi_{\hat{\theta}(s)} g_{\hat{\theta}(s)}\|_V$ is bounded.

This ends the proof of inequality (4.4) of **(A3'(i))**.

We now move to the Hölder conditions (4.5) of **(A3'(i))**. We will use the two following lemmas which state Hölder conditions on the transition kernel and its iterates:

Lemma 1 *Let \mathcal{K} be a compact subset of \mathcal{S} . There exists a constant $C_{\mathcal{K}}$ such that for all $1 \leq p$ there exists $q > p$, for all function $f \in \mathcal{L}_{V^p}$ and for all $(s, s') \in \mathcal{K}^2$ we have :*

$$\|\Pi_{\hat{\theta}(s)} f - \Pi_{\hat{\theta}(s')} f\|_{V^q} \leq C_{\mathcal{K}} \|f\|_{V^p} \|s - s'\|.$$

Proof For any $f \in \mathcal{L}_{V^p}$ and any $x \in \mathbb{R}^l$, we have

$$\begin{aligned} \Pi_s f(x) &= \int_{\mathbb{R}^l} f(z) \alpha_s(x, z) q_s(x, z) dz \\ &\quad + f(x)(1 - \alpha_s(x)), \end{aligned}$$

where $\alpha_s(x, z) = \min\left(1, \frac{p_{\hat{\theta}(s)}(z)q_s(z, x)}{q_s(x, z)p_{\hat{\theta}(s)}(x)}\right)$ and $\alpha_s(x) = \int \alpha_s(x, z) q_s(x, z) dz$ is the average acceptance rate. Let us denote for all x, z and s : $r_s(x, z) = \alpha_s(x, z) q_s(x, z)$.

Let s and s' be two points in \mathcal{K} . We note that $s \mapsto \hat{\theta}(s)$ is a continuously differentiable function therefore uniformly bounded in $s \in \mathcal{K}$.

$$\begin{aligned} \|\Pi_s f(x) - \Pi_{s'} f(x)\| &\leq \|f\|_{V^p} \times \\ &\left\{ \int_{\mathcal{X}} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz + \right. \\ &\left. V^p(x) \int_{\mathcal{X}} |r_s(x, z) - r_{s'}(x, z)| dz \right\}, \\ &\leq 2\|f\|_{V^p} V^p(x) \times \\ &\int_{\mathcal{X}} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz. \end{aligned}$$

Let $I = \int_{\mathcal{X}} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz$. For sake of simplicity, we denote by A_s the acceptance set instead of $A_s(x)$. We decompose I into four terms :

$$I = \int_{A_s \cap A_{s'}} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz \quad (51)$$

$$+ \int_{A_s \cap A_{s'}^c} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz \quad (52)$$

$$+ \int_{A_s^c \cap A_{s'}} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz \quad (53)$$

$$+ \int_{A_s^c \cap A_{s'}^c} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz \quad (54)$$

Let us first consider the term (51).

$$\begin{aligned} &\int_{A_s \cap A_{s'}} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz \\ &= \int_{A_s \cap A_{s'}} |q_s(x, z) - q_{s'}(x, z)| V^p(z) dz. \end{aligned} \quad (55)$$

We use the mean value theorem on the smooth function $s \mapsto q_s(x, z)$ for fixed values of (x, z) .

$$\frac{dq_s(x, z)}{ds} = q_s(x, z) \frac{d \log q_s(x, z)}{ds}.$$

After some calculations, using the bounded drift and covariance and Assumption **(M8)**, we get :

$$\begin{aligned} \frac{d \log q_s(x, z)}{ds} &\leq \tilde{P}_1(x, z) \left(\left\| \frac{dD_s(x)}{ds} \right\| + \left\| \frac{d\Sigma_s(x)}{ds} \right\|_F \right. \\ &\quad \left. + \left\| \frac{d\Sigma_s^{-1}(x)}{ds} \right\|_F \right), \\ &\leq C_{\mathcal{K}} P_1(x, z). \end{aligned}$$

where D_s and Σ_s are respectively the drift and covariance of the proposal q_s and \tilde{P}_1 and P_1 are two polynomial functions in both variables.

Using Equation (23), we have :

$$\left| \frac{dq_s(x, z)}{ds} \right| \leq k_2 C_{\mathcal{K}} P_1(x, z) g_{\epsilon_2}(z - x),$$

which leads to :

$$\begin{aligned} &\int_{A_s \cap A_{s'}} |q_s(x, z) - q_{s'}(x, z)| V^p(z) dz \\ &\leq k_2 C_{\mathcal{K}} \|s - s'\| \int_{\mathcal{X}} V^p(z) P_1(x, z) g_{\epsilon_2}(z - x) dz \\ &\leq k_2 C_{\mathcal{K}} Q_1(x) \|s - s'\|, \end{aligned}$$

where Q_1 is a polynomial function.

Now we move to the second term (52). Let $z \in A_s \cap A_{s'}^c$. We define for all $u \in [0, 1]$ the barycenter $s(u)$ of s and s' equals to $us + (1 - u)s'$ which belongs to the convex hull of the compact subset \mathcal{K} .

Since $u \mapsto \rho_{s(u)}(x, z)$ is continuously differentiable, $\rho_s(x, z) \geq 1$ and $\rho_{s'}(x, z) < 1$, using the intermediate value theorem, there exists $u \in]0, 1[$ depending on x and z such that $\rho_{s(u)}(x, z) = 1$. We choose the minimum value of u satisfying this condition. Therefore,

$$\begin{aligned} &|\alpha_s(x, z) q_s(x, z) - \alpha_{s'}(x, z) q_{s'}(x, z)| \\ &\leq |q_s(x, z) - q_{s(u)}(x, z)| + \\ &\quad |\rho_{s(u)}(x, z) q_{s(u)}(x, z) - \rho_{s'}(x, z) q_{s'}(x, z)|. \end{aligned} \quad (56)$$

We treat the first term of the right hand side as previously. For the second term, we use the mean value theorem for the function $v \mapsto f_{s(v)}(x, z) = \rho_{s(v)}(x, z) q_{s(v)}(x, z)$ on $]0, u[$. There exists $v \in]0, u[$ such that

$$|f_{s(u)}(x, z) - f_{s'}(x, z)| \leq \left| \frac{df_{s(v)}(x, z)}{dv} \right| \|s - s'\|.$$

Thanks to the upper bound above we get

$$\frac{d \log f_{s(v)}(x, z)}{dv} \leq C_{\mathcal{K}} P_2(x, z),$$

where P_2 is a polynomial function in both variables. Since on the segment defined by $s(u)$ and s' we have $\rho_s(x, z) \leq 1$:

$$\begin{aligned} \frac{df_{s(v)}(x, z)}{dv} &= f_{s(v)}(x, z) \frac{d \log f_{s(v)}(x, z)}{dv} \\ &\leq C_{\mathcal{K}} q_{s(v)}(x, z) P_2(x, z) \\ &\leq k_2 C_{\mathcal{K}} \|s - s'\| P_2(x, z) g_{\varepsilon_2}(z - x). \end{aligned}$$

This yields :

$$\begin{aligned} \int_{A_s \cap A_{s'}^c} |r_s(x, z) - r_{s'}(x, z)| V^p(z) dz \\ \leq k_2 C_{\mathcal{K}} \left(Q_1(x) + Q_2(x) \right) \|s - s'\|. \end{aligned}$$

The third term (53) is the symmetric one of the second.

Let us end with the last term (54).

$$\begin{aligned} \int_{A_s^c \cap A_{s'}} |\alpha_s(x, z) q_s(x, z) - \alpha_{s'}(x, z) q_{s'}(x, z)| \times \\ V^p(z) dz \\ = \int_{A_s^c \cap A_{s'}} |\rho_s(x, z) q_s(x, z) - \rho_{s'}(x, z) q_{s'}(x, z)| \times \\ V^p(z) dz. \end{aligned}$$

If for all $u \in]0, 1[$, $\rho_{s(u)}(x, y) < 1$ then this term can be treated as the second term of Equation (56). If there exists $u \in]0, 1[$ such that $\rho_{s(u)}(x, y) \geq 1$, we define u_0 and u_1 respectively the smallest and biggest elements in $]0, 1[$ such that $\rho_{s(u_0)} = \rho_{s(u_1)} = 1$. The first and last terms are treated as the previous case and the middle term is treated as the term (51).

Putting all these upper bounds together yields :

$$\| \Pi_s f(x) - \Pi_{s'} f(x) \| \leq 2 \|f\|_{V^p} V^p(x) Q(x) \|s - s'\|, \quad (57)$$

where Q is a polynomial function in $x \in \mathcal{X}$. Therefore, there exists a constant $q > p$ such that $V^p(x) Q(x) \leq V^q(x)$ which concludes the proof.

Lemma 2 *Let \mathcal{K} be a compact subset of \mathcal{S} . There exists a constant $C_{\mathcal{K}}$ such that for all $1 \leq p < q$, for all function $f \in \mathcal{L}_{V^p}$, for all $(s, s') \in \mathcal{K}^2$ and for all $k \geq 0$, we have:*

$$\| \Pi_{\hat{\theta}(s)}^k f - \Pi_{\hat{\theta}(s')}^k f \|_{V^q} \leq C_{\mathcal{K}} \|f\|_{V^p} \|s - s'\|.$$

Proof The proof of lemma 2 follows the line of the proof of Proposition B.2 of [6].

Thanks to the proofs of [5], we get that h is a Hölder function for any $0 < a < 1$ which leads to **(A3''(i))**.

We finally focus on the proof of **(A3''(ii))**.

Lemma 3 *Let \mathcal{K} be a compact subset of \mathcal{S} and $p \geq 1$. For all sequences $\gamma = (\gamma_k)_{k \geq 0}$ and $\varepsilon = (\varepsilon_k)_{k \geq 0}$ satisfying $\varepsilon_k < \bar{\varepsilon}$ for some $\bar{\varepsilon}$ sufficiently small, there exists $C_{\mathcal{K}} > 0$, such that for any $z_0 \in \mathcal{X}$, we have*

$$\sup_{s \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}_{z, s}^{\gamma} [V^p(z_k) \mathbf{1}_{\sigma(\mathcal{K}) \wedge \nu(\varepsilon) \geq k}] \leq C_{\mathcal{K}} V^p(z_0),$$

where $\mathbb{E}_{z, s}^{\gamma}$ is the expectation related to the non-homogeneous Markov chain $((z_k, s_k))$ started from (z, s) with step size sequence γ .

Proof Let K be a compact subset of Θ such that $\hat{\theta}(K) \subset K$. We note in the sequel, $\theta_k = \hat{\theta}(s_k)$. We have for $k \geq 2$, using the Markov property and the drift property (18) for V^p ,

$$\begin{aligned} \mathbb{E}_{z, s}^{\gamma} [V^p(z_k) \mathbf{1}_{\sigma(\mathcal{K}) \wedge \nu(\varepsilon) \geq k}] &\leq \mathbb{E}_{z, s}^{\gamma} [\Pi_{\theta_{k-1}} V^p(z_{k-1})] \\ &\leq \lambda \mathbb{E}_{z, s}^{\gamma} [V^p(z_{k-1})] + C. \end{aligned} \quad (58)$$

$$(59)$$

Iterating the same arguments recursively leads to :

$$\begin{aligned} \mathbb{E}_{z, s}^{\gamma} [V^p(z_k) \mathbf{1}_{\sigma(\mathcal{K}) \wedge \nu(\varepsilon) \geq k}] \\ \leq \lambda^k \mathbb{E}_{z, s}^{\gamma} [V^p(z_0)] + C \sum_{l=0}^{k-1} \lambda^l. \end{aligned}$$

Since $\lambda < 1$ and $V(z) \geq 1$ for all $z \in \mathcal{X}$, for all $k \in \mathbb{N}$, we have :

$$\begin{aligned} \mathbb{E}_{z, s}^{\gamma} [V^p(z_k) \mathbf{1}_{\sigma(\mathcal{K}) \wedge \nu(\varepsilon) \geq k}] &\leq V^p(z_0) + C \sum_{l=0}^{k-1} \lambda^l \\ &\leq V^p(z_0) \left(1 + \frac{C}{1 - \lambda} \right). \end{aligned}$$

This yields **(A3'(ii))** which concludes the proof of Theorem 1.

7.3 Proof of the Central Limit Theorem for the Estimated Sequence

The proof of Theorem 2 follows the lines of the proof of Theorem 25 of [12]. This theorem is an application of Theorem 24 of [12] in the case of Markovian dynamics. However, some assumptions required in Theorem 24 are not fulfilled in our case: the A-stability of the algorithm and the boundedness in infinite norm of the solution of the Poisson equation.

Consider the stochastic approximation:

$$s_k = s_{k-1} + \gamma_k h(s_{k-1}) + \gamma_k \eta_k, \quad (60)$$

where the remainder term is decomposed as follows:

$$\eta_k = \xi_k + \nu_k - \nu_{k-1} + r_k \quad (61)$$

with

$$\xi_k = g_{\hat{\theta}(s_{k-1})}(z_k) - \Pi_{\hat{\theta}(s_{k-1})} g_{\hat{\theta}(s_{k-1})}(z_{k-1}) \quad (62)$$

$$\nu_k = -\Pi_{\hat{\theta}(s_k)} g_{\hat{\theta}(s_k)}(z_k) \quad (63)$$

$$r_k = \Pi_{\hat{\theta}(s_k)} g_{\hat{\theta}(s_k)}(z_k) - \Pi_{\hat{\theta}(s_{k-1})} g_{\hat{\theta}(s_{k-1})}(z_k) \quad (64)$$

and for any $s \in \mathcal{S}$, $g_{\hat{\theta}(s)}$ is a solution of the Poisson equation $g - \Pi_{\hat{\theta}(s)} g = H_s - p_{\hat{\theta}(s)}(H_s)$.

We recall Theorem 24 of [12] with sufficient assumptions for our setting.

Theorem 3 (Adapted from Theorem 24 [12]) *Let assumptions (N1) and (N3) be fulfilled. Furthermore, assume that for some matrix U , some $\varepsilon > 0$ and some positive random variables X, X', X'' :*

The sequence (ξ_i) is a \mathcal{F} -martingale (65)

$$\sup_{i \in \mathbb{N}} \|\xi_i\|_{2+\varepsilon} < \infty \quad (66)$$

$$\lim_{k \rightarrow \infty} \gamma_k^{-1/2} \|X r_k\|_1 = 0 \quad (67)$$

$$\lim_{k \rightarrow \infty} \gamma_k^{1/2} \|X' \nu_k\|_1 = 0 \quad (68)$$

$$\lim_{k \rightarrow \infty} \gamma_k \|X'' \sum_{i=1}^k (\xi_i \xi_i^T - U)\|_1 = 0 \quad (69)$$

where $\mathcal{F} = (\mathcal{F}_i)_{i \in \mathbb{N}}$ is the increasing family of σ -algebra generated by the random variables (s_0, z_1, \dots, z_i) . Then

$$\frac{s_k - s^*}{\sqrt{\gamma_k}} \rightarrow_{\mathcal{L}} \mathcal{N}(0, V) \quad (70)$$

where V is the solution of the following Lyapunov equation $U + JV + VJ^T = 0$.

The result of Theorem 24 still holds replacing assumption (C) of [12] by (N1). Indeed, it is sufficient to establish that the random variable $\gamma_k^{-1/2} \sum_{i=0}^k \exp[(t_k - t_i)J] \gamma_i r_i$ converges toward 0 in probability where $t_i = \sum_{j=1}^i \gamma_j$. Theorem 19 and Proposition 39 of [12] can be applied in expectation. Theorems 23 and 20 of [12] also still hold with assumption (N1).

We now prove that assumptions of Theorem 3 hold.

By definition of ξ_i it is obvious that (65) is fulfilled. Moreover, the following lemma proves that there exists $\varepsilon > 0$ such that (66) holds with $X = 1$.

Lemma 4 *For all $\varepsilon > 0$, the sequence (ξ_k) is in $L^{2+\varepsilon}$.*

Proof We use the convexity of the function $x \mapsto x^{2+\varepsilon}$. Indeed, we have

$$\begin{aligned} & |g_{\hat{\theta}(s_{k-1})}(z_k) - \Pi_{\hat{\theta}(s_{k-1})} g_{\hat{\theta}(s_{k-1})}(z_{k-1})|^{2+\varepsilon} \\ & \leq (|g_{\hat{\theta}(s_{k-1})}(z_k)| + |\Pi_{\hat{\theta}(s_{k-1})} g_{\hat{\theta}(s_{k-1})}(z_{k-1})|)^{2+\varepsilon} \\ & \leq C_\varepsilon (|g_{\hat{\theta}(s_{k-1})}(z_k)|^{2+\varepsilon} + |\Pi_{\hat{\theta}(s_{k-1})} g_{\hat{\theta}(s_{k-1})}(z_{k-1})|^{2+\varepsilon}), \end{aligned}$$

where $C_\varepsilon = \frac{1}{2^{3+\varepsilon}}$.

Applying the drift condition, we get :

$$\begin{aligned} \mathbb{E}(\|\xi_k\|^{2+\varepsilon} | \mathcal{F}_{k-1}) & \leq C_\varepsilon \left(\mathbb{E}(|g_{\hat{\theta}(s_{k-1})}(z_k)|^{2+\varepsilon} | \mathcal{F}_{k-1}) \right. \\ & \quad \left. + \mathbb{E}(|\Pi_{\hat{\theta}(s_{k-1})} g_{\hat{\theta}(s_{k-1})}(z_{k-1})|^{2+\varepsilon} | \mathcal{F}_{k-1}) \right) \\ & \leq C \mathbb{E}(V(z_k)^{2+\varepsilon} + V(z_{k-1})^{2+\varepsilon} | \mathcal{F}_{k-1}) \\ & \leq C (\lambda V^{2+\varepsilon}(z_{k-1}) + 1). \end{aligned}$$

Finally taking the expectation after induction as in Lemma 3 leads to:

$$\mathbb{E}(\|\xi_k^{2+\varepsilon}\|) \leq C V^{2+\varepsilon}(z_0) < +\infty.$$

Let us now focus on Equation (67). Thanks to the Hölder property of our kernel and the fact that H_{s_k} belongs to \mathcal{L}_V :

$$\begin{aligned} \|r_k\|_1 & = \mathbb{E} \|\Pi_{\hat{\theta}(s_k)} g_{\hat{\theta}(s_k)}(z_k) - \Pi_{\hat{\theta}(s_{k-1})} g_{\hat{\theta}(s_{k-1})}(z_k)\| \\ & \leq C \mathbb{E}[V^q(z_k) | s_k - s_{k-1}|^q] \\ & \leq C \mathbb{E}[V^{q+1}(z_k)] \gamma_k^q \\ & \leq C \gamma_k^q \end{aligned}$$

where the last inequality comes from the drift property. Since the Hölder property is true for

any $0 < a < 1$, we can choose $a > 1/2$ which leads to the conclusion.

To prove Equation (68), we note that using the drift condition as in the previous lemma, $\mathbb{E}(\|\nu_k\|)$ is uniformly bounded in k . Since the step-size sequence $(\gamma_k)_k$ tends to zero, the result follows with $X' = 1$.

We follow the lines of the proof of Theorem 25 of [12] to establish Equation (69). As in the proof of Lemma 4, we use the drift property coupled with our Hölder condition in \mathcal{L}_V -norm instead of the usual Hölder condition considered by [12] which is denoted (MS).

This concludes the proof of Theorem 3.

Applying Theorem 3 allows to prove the first part of Theorem 2. Assumption **(N2)** enables to characterize the covariance matrix Γ . The Delta method gives the result on the sequence (θ_k) achieving the proof of our Central Limit Theorem.

References

- Aldridge, G., Ratnanather, J., Martone, M., Terada, M., Beg, M., Fong, L., Ceyhan, E., Kolasny, A., Brown, T., Cochran, E., Tang, S., Pisano, D., Vaillant, M., Hurdal, M., Churchill, J., Greenough, W., Miller, M., Ellisman, M.: Semi-automated shape analysis of dendrite spines from animal models of fragileX and parkinson's disease using large deformation diffeomorphic metric mapping. Society for Neuroscience Annual Meeting, Washington DC (2005)
- Allasonnière, S., Amit, Y., Trouvé, A.: Toward a coherent statistical framework for dense deformable template estimation. *JRSS* **69**, 3–29 (2007)
- Allasonnière, S., Kuhn, E.: Stochastic algorithm for bayesian mixture effect template estimation. *ESAIM Probab Stat* **14**, 382–408 (2010)
- Allasonnière, S., Kuhn, E., Trouvé, A.: Bayesian consistent estimation in deformable models using stochastic algorithms: Applications to medical images. *Journal de la Société Française de Statistique* **151(1)**, 1–16 (2010)
- Allasonnière, S., Kuhn, E., Trouvé, A.: Bayesian deformable models building via stochastic approximation algorithm: A convergence study. *Bernoulli J.* **16(3)**, 641–678 (2010)
- Andrieu, C., Moulines, E., Priouret, P.: Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44(1)**, 283–312 (electronic) (2005)
- Atchadé, Y.: An adaptive version for the metropolis adjusted langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.* **8**, 235–254 (2006)
- Bigot, J., Charlier, B.: On the consistency of Fréchet means in deformable models for curve and image analysis. *Electron. J. Stat.* **5**, 1054–1089 (2011). DOI 10.1214/11-EJS633. URL <http://dx.doi.org/10.1214/11-EJS633>
- Ceyhan, E., Fong, L., Tasky, T., Hurdal, M., Beg M.F. and Martone, M., Ratnanather, J.: Type-specific analysis of morphometry of dendrite spines of mice. 5th Int. Symp. Image Signal Proc. Analysis, ISPA pp. 7–12 (2007)
- Ceyhan, E., Ölken, R., Fong, L., Tasky, T., Hurdal, M., Beg, M., Martone, M., Ratnanather, J.: Modeling metric distances of dendrite spines of mice based on morphometric measures. *Int. Symp on Health Informatics and Bioinformatics* (2007)
- Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models: their training and application. *Comp. Vis. and Image Understanding* **61(1)**, 38–59 (1995)
- Delyon, B.: Stochastic approximation with decreasing gain: convergence and asymptotic theory. Technical Report: Publication interne 952, IRISA (2000)
- Delyon, B., Lavielle, M., Moulines, E.: Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27(1)**, 94–128 (1999)
- Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **1**, 1–22 (1977)
- Gilks, W., Richardson, S., Spiegelhalter, D.: Markov Chain Monte Carlo in Practice. Chapman & Hall (1996)
- Girolami, M., Calderhead, B.: Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B* **73(2)**, 1–37 (2011)
- Glasbey, C.A., Mardia, K.V.: A penalised likelihood approach to image warping. *Journal of the Royal Statistical Society, Series B* **63**, 465–492 (2001)
- Grenander, U., Miller, M.I.: Computational anatomy: An emerging discipline. *Quarterly of Applied Mathematics* **LVI(4)**, 617–694 (1998)
- Jarner, S., Hansen, E.: Geometric ergodicity of metropolis algorithms. In: *Stochastic Processes and Their Applications*, pp. 341–361 (1998)
- Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage* **23**, S151–S160 (2004)
- Kuhn, E., Lavielle, M.: Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.* **8**, 115–131 (electronic) (2004)
- Maire, F., Lefebvre, S., Moulines, E., Douc, R.: Aircraft classification with low infrared sensor. *Statistical Signal Processing Workshop (SSP), IEEE* (2011)
- Marshall, T., Roberts, G.: An adaptive approach to langevin MCMC. *Statistics and Computing* **22 (5)**, 1041–1057 (2012)

24. Marsland, S., Twining, C.: Constructing diffeomorphic representations for the groupwise analysis of non-rigid registrations of medical images. *IEEE Transactions on Medical Imaging* **23** (2004)
25. Meyn, S.P., Tweedie, R.L.: Markov chains and stochastic stability. Communications and Control Engineering Series. Springer-Verlag London Ltd., London (1993)
26. Micheli, M., Michor, P.W., Mumford, D.B.: Sectional curvature in terms of the cometric, with applications to the riemannian manifolds of landmarks. *SIAM Journal on Imaging Sciences* **5**(1), 394–433 (2012)
27. Miller, M., Priebe, C., Qiu, A., Fischl, B., Kolasny, A., Brown, T., Park, Y., Ratnanather, J., Busa, E., Jovicich, J., Yu, P., Dickerson, B., Buckner, R.: Morphometry BIRN. collaborative computational anatomy: An MRI morphometry study of the human brain via diffeomorphic metric mapping. *Human Brain Mapping* **30**(7), 2132–2141 (2009)
28. Miller, M.I., Trouvé, A., Younes, L.: On the metrics and Euler-Lagrange equations of computational anatomy. *Annual Review of biomedical Engineering* **4** (2002)
29. Richard, F., Samson, A., Cuenod, C.A.: A SAEM algorithm for the estimation of template and deformation parameters in medical image sequences. *Statistics and Computing* **19**, 465–478 (2009)
30. Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2**(4), 341–363 (1996). DOI 10.2307/3318418. URL <http://dx.doi.org/10.2307/3318418>
31. Stramer, O., Tweedie, R.: Langevin-type models i: Diffusions with given stationary distributions, and their discretizations. *Methodol. Comput. Appl. Probab.* **1**(3), 283–306 (1999)
32. Stramer, O., Tweedie, R.: Langevin-type models ii: self-targeting candidates for mcmc algorithms. *Methodol. Comput. Appl. Probab.* **1**(3), 307–328 (1999)
33. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: Efficient non-parametric image registration. *Neuroimage* **45**, 61–72 (2009)