BAYESIAN DEFORMABLE MODELS BUILDING VIA STOCHASTIC APPROXIMATION ALGORITHM: A CONVERGENCE STUDY

S. ALLASSONNIÈRE , E. KUHN*, AND A. TROUVÉ[†]

Abstract. The problem of the definition and the estimation of generative models based on deformable templates from raw data is of particular importance for modelling non aligned data affected by various types of geometrical variability. This is especially true in shape modelling in the computer vision community or in probabilistic atlas building for Computational Anatomy (CA). A first coherent statistical framework modelling the geometrical variability as hidden variables has been given by Allassonnière, Amit and Trouvé in [1]. Setting the problem in a Bayesian context they proved the consistency of the MAP estimator and provided a simple iterative deterministic algorithm with an EM flavour leading to some reasonable approximations of the MAP estimator under low noise conditions. In this paper we present a stochastic algorithm for approximating the MAP estimator in the spirit of the SAEM algorithm. We prove its convergence to a critical point of the observed likelihood with an illustration on images of handwritten digits.

Key words. stochastic approximation algorithms, non rigid-deformable templates, shapes statistics, Bayesian modelling, MAP estimation.

AMS subject classifications. 60J22, 62F10, 62F15, 62M40.

1. Introduction.

The statistical analysis of high dimensional data is one of the most active fields in modern statistics nowadays. However, despite huge progress in the general theory of non-parametric statistics or machine learning, the practical efficiency of many "black box" universal methods can be quite limited if the invariances and structural constraints of a specific field are not properly taken into account. In particular, in the field of image analysis, the statistical analysis and modelling of variable objects from a limited set of examples is still a quite challenging and a largely unsolved problem. The analysis of shape variability, even coded as functional data thanks to the imaging process cannot be efficiently done "as is" without using a more adequate representation. One such representation is the so-called dense deformable template framework [2], where the observations are defined as deformations of a given exemplar or template under a family of deformations of moderate "dimensionality". Such a representation appears particularly adapted in the context of probabilistic atlases in Computational Anatomy where one aims at building a statistical model of the variability of anatomical data among a given population [8]. Whereas the statistical shape analysis theory based on a finite dimensional coding of shapes by landmarks is well developed [6], the dense deformable templates is more complex and challenging. Until recently, dense deformable templates have been studied mainly from a variational point of view as an efficient vehicle for a large range of registration algorithms [4], but the study of deformable templates from a statistical point of view as a class of generative models for images of deformable objects is still widely open. A major issue is the design of statistically sound algorithms for the estimation of dense deformable models from a sample of images of moderate size. A first approach in this direction were proposed in [7] or more recently in [10], the first used a penalised likelihood approach and the second one an MDL approach to estimate the template from a training set of non-aligned images. However, in both cases, the framework does not really differ from the variational approach in particular because the deformations are considered as *nuisance parameters* which need to be estimated. In consequence, the associated algorithms do not lead to consistent estimators of the template for generative models.

In this paper, we consider the hierarchical Bayesian framework for dense deformable templates developed by Allassonnière, Amit and Trouvé in [1] where each image in a given population is assumed to be generated as a noisy and randomly deformed version of a common template drawn from a prior distribution on the set of templates. The individual deformations appear as *hidden variables* (or random effects in the mixed effects terminology) whereas the template and the law of

*LAGA, Université Paris 13, 99, Av. Jean-Baptiste Clément, F-93430 Villetaneuse, France

[†]CMLA, ENS Cachan, CNRS, PRES UniverSud, 61 Av. Président Wilson, F-94230 Cachan, France

the deformations are parameters of interest for the estimation problem (or fixed effects). In [1] the estimation of the parameters (template and geometric deformation law) is performed by Maximum A Posteriori (MAP) and the existence and consistence of the MAP estimator is proved. On the algorithmic side, a deterministic iterative method, based on EM, is proposed to compute the MAP estimator. Nevertheless, the E step which consists in computing an expectation with respect to the a posteriori density is untractable in the current framework and a simple approximation by the mode of the posterior is proposed. This reduces to a registration problem of the current template to the images in the sample with a regularisation term given by the log-likelihood of the current deformation law. The result is a purely deterministic algorithm, alternating registration steps with updates of the template and of the geometric deformation law, and derived from a coherent statistical perspective. However, due to the approximation of the posterior by its mode, the convergence of the algorithm to the MAP does not hold even if it produces good results under low noise conditions.

Our goal in this paper is to overcome the limitations of this deterministic method as exhibited by several experiments and to propose a stochastic iterative method to compute the MAP estimator for which we will be able to prove convergence results. The solution proposed is to use a stochastic approximation of the EM algorithm: the non observed variables will be simulated. In the one component case (pure deformable model, no mixture) introduced by [1], we use the stochastic approximation EM (SAEM) algorithm coupled with a Markov Chain Monte Carlo method introduced by Kuhn and Lavielle in [9].

This algorithm has been proved to be convergent under the assumption, among others, that the non observed variables live in a compact set. This is not the case in our framework so we adapt this algorithm and also the convergence proof to a non compact setting by introducing truncation on random boundaries along the lines of [3].

The paper is organised as follows: in Section 2 we first recall the observation model proposed by Allassonnière, Amit and Trouvé in [1]. Then we describe in Section 3 the stochastic algorithm proposed in the one component case and give a convergence theorem. Section 4 is devoted to the experiments. To prove the convergence of our stochastic algorithm for deformable template estimation, we first state in Section 5 a rather general stability result for truncated stochastic approximation algorithms adapted from [3] and we show in Section 6 that it applies to MAP based deformable template estimation.

2. Observation model.

Let us recall the model introduced in [1]. We are given gray level images $(y_i)_{1\leq i\leq n}$ observed on a grid of pixels $\{r_s \in D \subset \mathbb{R}^2, s\in\Lambda\}$ where D is a continuous domain and Λ the pixel network. Although the images are observed only at the pixels $(r_s)_s$ we are looking for a template image $I_0 : \mathbb{R}^2 \to \mathbb{R}$ defined on the plane (the extension to images on $\mathbb{R}^d$ is straightforward). For each observation y, we assume the existence of an unobserved deformation field $z : \mathbb{R}^2 \to \mathbb{R}^2$ such that for $s \in \Lambda$

$$y(s) = I_0(r_s - z(r_s)) + \sigma\epsilon(s)$$

where $\sigma\epsilon$ denotes an additive noise.

2.1. Models for template and deformation.

Our model takes into account two complementary sides: photometry -indexed by p, and geometry -indexed by g. The template I_0 and the deformation z are assumed to belong to reproducing kernel Hilbert spaces V_p and V_g defined by their respective kernels K_p and K_g. Moreover we restrict them to the subset of linear combinations of the kernels centred at some fixed control points in the domain D: $(r_{p,k})_{1\leq k\leq k_p}$ respectively $(r_{g,k})_{1\leq k\leq k_g}$. They are therefore parametrised by the coefficients $\alpha \in \mathbb{R}^{k_p}$ and $\beta = (\beta^{(1)}, \beta^{(2)}) \in \mathbb{R}^{k_g} \times \mathbb{R}^{k_g}$ which yield to: $\forall r \in D$,

$$I_\alpha(r) = (\mathbf{K_p}\alpha)(r) = \sum_{k=1}^{k_p} K_p(r, r_{p,k})\alpha(k) \tag{2.1}$$

and

$$z_\beta(r) = (\mathbf{K_g}\beta)(r) = \sum_{k=1}^{k_g} K_g(r, r_{g,k})(\beta^{(1)}(k), \beta^{(2)}(k)). \tag{2.2}$$

Other forms of smooth parametric representations of the images and of the deformation fields could be used without changing the overall results.

2.2. Parametric model. We suppose that all the data can be explained through that statistical model (we denote below $y_1^n = (y_i)_{1\leq i\leq n}$ and $\beta_1^n = (\beta_i)_{1\leq i\leq n}$):

$$\begin{cases} \beta_1^n \sim \otimes_{i=1}^n \mathcal{N}_{2k_g}(0, \Gamma_g) \mid \Gamma_g \\ y_1^n \sim \otimes_{i=1}^n \mathcal{N}_{|\Lambda|}(z_{\beta_i} I_\alpha, \sigma^2 \mathrm{Id}) \mid \beta_1^n, \alpha, \sigma^2 \end{cases} \tag{2.3}$$

where $zI_\alpha(s) = I_\alpha(r_s - z(r_s))$, for s in Λ. The parameters of interest are α, σ^2 - the variance of the additive noise - and the covariance matrix Γ_g of the variables β. We assume that $\theta = (\alpha, \sigma^2, \Gamma_g)$ belongs to the parameter space Θ defined as the open set

$$\Theta \triangleq \{ \ \theta = (\alpha, \sigma^2, \Gamma_g) \mid \alpha \in \mathbb{R}^{k_p}, \ |, \ \sigma > 0, \ \Gamma_g \in \mathrm{Sym}^+_{2k_g} \ \},$$

where $\mathrm{Sym}^+_{2k_g}$ is the cone of real positive $2k_g \times 2k_g$ definite symmetric matrices.

The likelihood of the observed data can be written as an integral over the unobserved deformation parameters:

$$q(y_1^n|\theta) = \int q(y_1^n|\beta_1^n, \alpha, \sigma^2) q(\beta_1^n|\Gamma_g) d\beta_1^n$$

where all the densities are determined by the model. We denote all density functions as q.

2.3. Bayesian model. Even though the parameters are finite dimensional, the maximum-likelihood estimator can yield degenerate estimates when the training sample is small. Introducing prior distributions on the parameters, estimation with small samples is still possible and their effect can be seen in the parameter update steps [1]. We use a generative model which includes standard conjugate prior distributions with fixed hyperparameters: a normal prior on α and inverse-Wishart priors on σ^2 and Γ_g. All priors are assumed independent: $\theta = (\alpha, \sigma^2, \Gamma_g) \sim \nu_p \otimes \nu_g$ where

$$\begin{cases} \nu_p(d\alpha, d\sigma^2) \propto \exp\left(-\frac{1}{2}(\alpha - \mu_p)^t (\Sigma_p)^{-1}(\alpha - \mu_p)\right) \left(\exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right) \frac{1}{\sqrt{\sigma^2}}\right)^{a_p} d\sigma^2 d\alpha, \ a_p \geq 3 \\ \nu_g(d\Gamma_g) \propto \left(\exp(-\langle \Gamma_g^{-1}, \Sigma_g\rangle_F/2) \frac{1}{\sqrt{|\Gamma_g|}}\right)^{a_g} d\Gamma_g, \ a_g \geq 4k_g + 1 \ . \end{cases} \tag{2.4}$$

For two matrices A, B we define $\langle A, B\rangle_F \triangleq tr(A^t B)$.

3. Parameters estimation with stochastic approximation EM. In our Bayesian framework, we obtain from [1] the existence of the MAP estimator

$$\hat{\theta}_n = \operatorname*{argmax}_\theta q(\theta|y_1^n) \, .$$

We turn now to the maximisation problem of the posterior distribution $q(\theta|y_1^n)$ and we recall here briefly how the stochastic approach can be derived from the computation of the derivative of the observed log-likelihood.

3.1. Formal derivation of the stochastic approach.

NOTATION 1. *To simplify the presentation, let us denote in the sequel $x \triangleq \beta_1^n \in \mathbb{R}^N$ with $N \triangleq 2nk_g$ the vector collecting all the missing variables and $y \triangleq y_1^n$ the collection of observations.*

Consider curved exponential densities, that is to say, situations where the complete likelihood can be written as:

$$q(y, x, \theta) = \exp\left[-\psi(\theta) + \langle S(x), \phi(\theta)\rangle\right] \tag{3.1}$$

where the sufficient statistic S is a Borel function on $\mathbb{R}^N$ taking its values in an open subset $\mathcal{S}$ of $\mathbb{R}^m$ and ψ, ϕ two Borel functions on Θ (note that S, ϕ and ψ may depend also on y, but since y will stay fixed in the sequel, we omit this dependency).

As we are looking for the MAP estimator, we try to maximise the observed log-likelihood $l(\theta) \triangleq \log \int q(y, x, \theta)dx$. One solution would be a steepest ascent method where the expression of the gradient is (we ignore the problem of existence of the derivatives):

$$\frac{\partial l}{\partial \theta}(\theta) = \mathbb{E}_\theta\left[S(X)^t \frac{\partial \phi}{\partial \theta}(\theta) \ \mid\ y\right] - \frac{\partial \psi}{\partial \theta}(\theta)\,,$$

where $\mathbb{E}_\theta\left[f(X)\,|\,y\right] \triangleq \int f(x)q(x|y, \theta)dx$ for any $q(x|y, \theta)dx$-integrable Borel mapping $x \to f(x)$. We introduce the following function: $L : \mathcal{S} \times \Theta \to \mathbb{R}$ as

$$L(s; \theta) = -\psi(\theta) + \langle s, \phi(\theta)\rangle \tag{3.2}$$

and suppose there exists a function $\hat{\theta} : \mathcal{S} \to \Theta$ such that:

$$\forall \theta \in \Theta, \forall s \in \mathcal{S}, L(s; \hat{\theta}(s)) \geq L(s; \theta)\ . \tag{3.3}$$

Then,

$$\frac{\partial L}{\partial \theta}(s, \hat{\theta}(s)) = s^t \frac{\partial \phi}{\partial \theta}(\hat{\theta}(s)) - \frac{\partial \psi}{\partial \theta}(\hat{\theta}(s)) = 0 \tag{3.4}$$

and

$$\frac{\partial l}{\partial \theta}(\hat{\theta}(s)) = \mathbb{E}_\theta\left[(S(X) - s)^t\ \frac{\partial \phi}{\partial \theta}(\hat{\theta}(s))\middle|\, y\right]\ . \tag{3.5}$$

Note that s becomes the natural variable, so that the gradient can be computed with respect to s. This yields:

$$\frac{\partial l \circ \hat{\theta}(s)}{\partial s} = \frac{\partial l}{\partial \theta}(\hat{\theta}(s))\frac{d\hat{\theta}}{ds}(s)\ .$$

From (3.4), $\frac{d}{ds}\left(\frac{\partial L}{\partial \theta}(s, \hat{\theta}(s))\right) = \frac{\partial^2 L}{\partial s \partial \theta}(s, \hat{\theta}(s)) + \frac{\partial^2 L}{\partial \theta^2}(s, \hat{\theta}(s))\frac{d\hat{\theta}}{ds}(s) = 0$ so that computing $\frac{\partial^2 L}{\partial s \partial \theta} = \frac{\partial^2 L}{\partial \theta \partial s}$ from equation (3.2), we get:

$$-\left(\frac{\partial^2 L}{\partial \theta^2}\right)(s, \hat{\theta}(s))\frac{d\hat{\theta}}{ds}(s) = \frac{\partial \phi}{\partial \theta}(\hat{\theta}(s))^t\ .$$

Finally, for $M(s) = -\left(\frac{d\hat{\theta}}{ds}(s)\right)^t \left(\frac{\partial^2 L}{\partial \theta^2}(s, \hat{\theta}(s))\right)\frac{d\hat{\theta}}{ds}(s)$, we have

$$\frac{\partial l \circ \hat{\theta}}{\partial s}(s) = \mathbb{E}_\theta\left[(S(X) - s)^t M(s)\,|\,y\right]\ .$$

Since $\hat{\theta}(s)$ is a maximum, $\left(\frac{\partial^2 L}{\partial \theta^2}\right)$ is symmetric non positive and $M(s)$ is a symmetric non negative matrix so that if

$$w(s) \triangleq -l \circ \hat{\theta}(s) \text{ and } h(s) \triangleq \mathbb{E}_{\hat{\theta}(s)}\left[(S(X) - s)\,|\,y\right]\ , \tag{3.6}$$

we have

$$\langle \frac{\partial l \circ \hat{\theta}}{\partial s}(s), h(s) \rangle = h(s)^t M(s) h(s) \geq 0 \tag{3.7}$$

and $h(s)$ is always a descent direction of w. Thus the ODE

$$\frac{ds}{dt} = h(s(t))$$

defines a trajectory for which $w(s(t))$ is decreasing i.e. $l \circ \hat{\theta}(s(t))$ is increasing. An Euler discretisation of the previous ODE leads to:

$$s_k - s_{k-1} = \Delta_k h(s_{k-1}) = \Delta_k \mathbb{E}_{\hat{\theta}(s_{k-1})}\left[S(X) - s_{k-1}|y\right]$$

where $(\Delta_k)_{k \geq 0}$ is the decreasing time-step sequence. As the expectation is intractable, the usual route is to use a simulation of the missing data x_k: For any $s \in \mathcal{S}$, let $H_s : \mathbb{R}^N \to \mathcal{S}$ such that

$$H_s(x) \triangleq S(x) - s\,, \tag{3.8}$$

we have

$$h(s) = \mathbb{E}_{\hat{\theta}(s)}\left[H_s(X)|y\right]$$

so that

$$s_k - s_{k-1} = \Delta_k h(s_{k-1}) \simeq \Delta_k H_{s_{k-1}}(x_k) = \Delta_k (S(x_k) - s_{k-1})\,. \tag{3.9}$$

Since x_k cannot be easily drawn from the posterior distribution $q(x|y, \hat{\theta}(s_{k-1}))$, the usual alternative in stochastic approximation of ODEs is to simulate x_k from x_{k-1} with a Markov kernel having $q(x|y, \hat{\theta}(s_{k-1}))$ as stationary distribution.

3.2. SAEM-MCMC algorithm with truncation on random boundaries. In fact, the stochastic algorithm derived in the previous section is nothing but the so called Stochastic Approximation EM coupled with a Monte Carlo Markov Chain procedure proposed by [9] which generalised the algorithm introduced by [5]. Indeed, the k^{th} iteration of the SAEM-MCMC algorithm consists of three steps:

Simulation step the missing data, here the deformation parameters $x = \beta_1^n$, are drawn using a transition probability of a convergent Markov Chain Π_θ having the posterior distribution $\pi_\theta = q(x|y, \theta)$ as stationary distribution,

$$x_k \sim \Pi_{\theta_{k-1}}(x_{k-1}, \cdot)\,.$$

Stochastic approximation step a stochastic approximation is done on the complete log-likelihood using the simulated value of the missing data,

$$Q_k(\theta) = Q_{k-1}(\theta) + \Delta_{k-1}[\log q(y, x_k, \theta) - Q_{k-1}(\theta)]$$

where $(\Delta_k)_k$ is a decreasing sequence of positive step-sizes.

Maximisation step the parameters are updated in the M-step,

$$\theta_k = \underset{\theta}{\operatorname{argmax}}\, Q_k(\theta)\,.$$

The initial values of Q and θ are arbitrary chosen.

Since our model is exponential, the stochastic approximation can be done on the complete log-likelihood as well as on a sufficient statistic. This is due to the fact that the missing data only appears linearly through a sufficient statistic S in the exponential exponent. This yields the following stochastic approximation:

$$s_k = s_{k-1} + \Delta_{k-1}(S(x_k) - s_{k-1}) \tag{3.10}$$

which is none other than equation (3.9) in the previous section.

However, as we set a Gaussian prior on the missing variables x, we cannot assume its support is compact as in [9]. We thus have to employ the more general setting introduced in [3] which involves truncation on random boundaries. Thanks to this approach, we end up with an algorithm using an MCMC coupling procedure on SAEM and the truncation on random boundaries detailed below.

Let $(\mathcal{K}_q)_{q\geq 0}$ be an increasing sequence of compact subsets of $\mathcal{S}$ such as $\cup_{q\geq 0}\mathcal{K}_q = \mathcal{S}$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1}), \forall q \geq 0$. Let $(\varepsilon_k)_{k\geq 0}$ be a monotone non-increasing sequence of positive numbers and K a subset of $\mathbb{R}^N$. We construct the sequence $((s_k, x_k))_{k\geq 0}$ as explained in Algorithm 1. As long as the stochastic approximation does not wander outside the current compact set and is not too far from its previous value, we run the SAEM-MCMC algorithm. As soon as one of the two previous conditions is not satisfied, we reinitialise the sequences of s and x using a projection (for more details see [3]).

Algorithm 1 Stochastic approximation with truncation on random boundaries

```
Set $\kappa_0 = 0$, $s_0 \in \mathcal{K}_0$ and $x_0 \in$ K.
for all $k \geq 1$ do
  compute $\bar{s} = s_{k-1} + \Delta_{k-1}(S(\bar{x}) - s_{k-1})$
  where $\bar{x}$ is sampled from a transition kernel $\Pi_{\theta_{k-1}}(x_{k-1}, .)$.
  if $\bar{s} \in \mathcal{K}_{\kappa_{k-1}}$ and $|\bar{s} - s_{k-1}| \leq \varepsilon_{k-1}$ then
    set $(s_k, x_k) = (\bar{s}, \bar{x})$ and $\kappa_k = \kappa_{k-1}$,
  else
    set $(s_k, x_k) = (\tilde{s}, \tilde{x}) \in \mathcal{K}_0 \times$ K and $\kappa_k = \kappa_{k-1} + 1$,
    where $(\tilde{s}, \tilde{x})$ can be chosen through different ways cf([3]).
  end if
  $\theta_k = \underset{\theta}{\text{argmax}}\, L(s_k, \theta)$
end for
```

3.3. Transition probability of the Markov Chain. We now explain how we simulate the Markov Chain of the missing variables $x = \beta_1^n$ given the observations $y = y_1^n$. The vector x is an element of the high dimensional space $\mathbb{R}^N$ and to face the potential problems due to this high dimensionality, we use a hybrid Gibbs sampler scanning all the coordinates x^j. For each j, let us denote $x_{-j} = (x^l)_{l\neq j}$. The coordinate x^j is not refreshed according to the usual conditional density $q(x^j|x_{-j}, y, \theta)$ which is not easily available but according to the Hasting Metropolis algorithm whose proposal law is given by $q(x^j|x_{-j}, \theta)$ i.e. the *a priori* conditional law according to the current parameter value θ.

For any $b \in \mathbb{R}$ and $1 \leq j \leq N$, denote by $x_{j,b}$ the unique configuration which is equal to x everywhere except in j where $x_{j,b}^j = b$. If b is proposed by the proposal law at coordinate j, the acceptance ratio is as usual given by $r_{\theta,j}(x, b) = \left[\frac{q(x_{j,b}|y,\theta)q(x^j|x_{-j},\theta)}{q(x|y,\theta)q(b|x_{-j},\theta)} \wedge 1\right]$. Since

$$q(x^j|x_{-j}, y, \theta) \propto q(y|x, \theta)q(x^j|x_{-j}, \theta)$$

the acceptance ratio can be simplified to

$$r_{\theta,j}(x, b) = \left[\frac{q(y|x_{j,b}, \theta)}{q(y|x, \theta)} \wedge 1\right] .$$

This hybrid Gibbs sampler is explained in [12] among others and Algorithm 2 summarises a transition step of the Markov Chain generation.

We denote $\Pi_{\theta,j}(x, dz) = q(z^j|x_{-j}, \theta)r_{\theta,j}(x, z^j)\mathbf{1}_{z_{-j}=x_{-j}}dz$ for $z^j \neq x^j$ the associated kernel on x defined by the update of the jth coordinate and $\Pi_\theta = \Pi_{\theta,N} \cdots \Pi_{\theta,1}$ the kernel associated with a complete scan.

Algorithm 2 Transition step $k \to k+1$ using a hybrid Gibbs sampler

```
Require: $x = x_k$; $\theta = \theta_k$
  Gibbs sampler:
  for all $j = 1 : N$ do
    Hasting-Metropolis procedure:
    $b \sim q(b|x_{-j}, \theta)$;
    Compute $r_{\theta,j}(x,b) = \left[\frac{q(y|x_{j,b},\theta)}{q(y|x,\theta)} \wedge 1\right]$
    With probability $r_{\theta,j}(x,b)$, update $x^j$: $x^j \leftarrow b$
  end for
```

3.4. Convergence Theorem. In this section, we give a convergence result for the truncated procedure in the case where the noise variance σ^2 *is fixed*, taking into account only the template α and the geometric covariance matrix Γ_g as our parameters (indeed, if σ is free we have not succeeded in providing a simple proof of the C^1 regularity of the mapping $s \to \hat{\theta}(s)$ which is usually needed in this setting). Hence, in this section $\theta = (\alpha, \Gamma_g)$.

We define the sufficient statistics in this setting. The complete log-likelihood can be written as:

$$\log q(y,x,\theta) = \log q(y|x,\theta) + \log q(x|\theta) + \log q(\theta) \ ,$$

so that, denoting $\forall 1 \leq k \leq k_p$ and $\forall s \in \Lambda$, the coordinate (s,k) of the matrix K_p^β

$$K_p^\beta(s,k) = K_p(r_s - z_\beta(r_s), r_{p,k}) \ ,$$

then

$$\begin{aligned}\log q(y,x,\theta) = &\sum_{i=1}^n \left\{ -\frac{|\Lambda|}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}|y_i - K_p^{\beta_i}\alpha|^2 \right\} \\ &+ \sum_{i=1}^n \left\{ -\frac{2k_g}{2}\log(2\pi) - \frac{1}{2}\log(|\Gamma_g|) - \frac{1}{2}\beta_i^t\Gamma_g^{-1}\beta_i \right\} \\ &+ a_g\left\{ -\frac{1}{2}\langle\Gamma_g^{-1}, \Sigma_g\rangle_F - \frac{1}{2}\log(|\Gamma_g|) \right\} - \frac{1}{2}(\alpha-\mu_p)^t\Sigma_p^{-1}(\alpha-\mu_p) \ .\end{aligned}$$

Developing the square $|y_i - K_p^{\beta_i}\alpha|^2$ and using the fact that $\langle K_p^{\beta_i}\alpha, K_p^{\beta_i}\alpha\rangle = \langle (K_p^{\beta_i})^t K_p^{\beta_i}, \alpha^t\alpha\rangle_F$, we get easily the following matricial form of the sufficient statistics:

$$S_1(x) = \sum_{1\leq i\leq n} \left(K_p^{\beta_i}\right)^t y_i \tag{3.11}$$

$$S_2(x) = \sum_{1\leq i\leq n} \left(K_p^{\beta_i}\right)^t \left(K_p^{\beta_i}\right) \tag{3.12}$$

$$S_3(x) = \sum_{1\leq i\leq n} \beta_i^t\beta_i \ . \tag{3.13}$$

For simplicity, we denote $S(x) = (S_1(x), S_2(x), S_3(x))$ for any $x = \beta_1^n \in \mathbb{R}^N$ and define the sufficient statistic space as

$$\mathcal{S} = \left\{ (s_1,s_2,s_3) | s_1 \in \mathbb{R}^{k_p}, \ s_2 + \sigma^2\Sigma_p^{-1} \in \mathrm{Sym}^+_{k_p}, \quad s_3 + a_g\Sigma_g \in \mathrm{Sym}^+_{2k_g} \right\} \ .$$

Identifying s_2 and s_3 with their lower triangular parts, the set $\mathcal{S}$ can be viewed as an open set of $\mathbb{R}^{n_s}$ with $n_s = k_p + \frac{k_p(k_p+1)}{2} + k_g(2k_g+1)$.

As already proved in [1] the maximising function $\hat{\theta}$ satisfying (3.3) exists. We can thus give an explicit form of $\hat{\theta}(s) = (\alpha(s), \Gamma_g(s))$ for our sufficient statistic vectors and matrices (s_1, s_2, s_3):

$$\begin{aligned}\Gamma_g(s) &= \tfrac{1}{n+a_g}(s_3 + a_g\Sigma_g)\ , \\ \alpha(s) &= \left(s_2 + \sigma^2(\Sigma_p)^{-1}\right)^{-1}\left(s_1 + \sigma^2(\Sigma_p)^{-1}\mu_p\right)\ .\end{aligned} \tag{3.14}$$

All these formula also prove the smoothness of $\hat{\theta}$ on the subset $\mathcal{S}$. This property enables us to work either with the stochastic approximation variable s or with the parameter function $\hat{\theta}(s)$ when needed in Algorithm 1.

As said before, the proof of the convergence of the stochastic sequence (s_k) to critical points of the observed log-likelihood in our model cannot rely on the coupling result given in [9] because of the restrictive assumption on the compactness of the missing data support (since we set a Gaussian prior on the missing variable β, we should not restrict the estimation of the law to any compact subset). We also cannot apply directly the convergence results proved in [3] about the stability of stochastic approximation since they assume several Hölder conditions involving the Markov transition kernel which are not fully satisfied in our model. However it is possible to adapt their proof while partially relaxing some of the assumptions and obtain the same convergence results. This technical part is postponed to Section 5 from which we can deduce our first result:

Theorem 3.1 (Convergence of Bayesian deformable template building via SA). *Let $w(s) = -l \circ \hat{\theta}(s)$ and $h(s) = \int (S(x) - s)q(x|y, \hat{\theta}(s))dx$ for $s \in \mathcal{S}$. Assume that:*

1. *there exist $p \geq 2$ and $a \in]0, 1[$ such that the sequences $\Delta = (\Delta_k)_{k\geq 0}$ and $\varepsilon = (\varepsilon_k)_{k\geq 0}$ are non-increasing, positive and satisfy:*
   $\sum\limits_{k=0}^{\infty} \Delta_k = \infty$, $\lim\limits_{k\to\infty} \varepsilon_k = 0$ *and* $\sum\limits_{k=1}^{\infty} \{\Delta_k^2 + \Delta_k \varepsilon_k^a + (\Delta_k \varepsilon_k^{-1})^p\} < \infty$;
2. $\mathcal{L}' \triangleq \{s \in \mathcal{S}, \langle \nabla w(s), h(s)\rangle = 0\}$ *is included in a level set of w.*

Let $(s_k)_{k\geq 0}$ be the sequence defined in Algorithm 1 with K *bounded and $\mathcal{K}_0 \subset S(\mathbb{R}^N)$. Then, for all $x_0 \in$ K and $s_0 \in \mathcal{K}_0$, we have*

$$\lim_{k\to\infty} d(s_k, \mathcal{L}') = 0 \ \bar{\mathbb{P}}_{x_0,s_0} \text{ -a.s.},$$

where $\bar{\mathbb{P}}_{x_0,s_0}$ is the probability measure associated with the chain $Z_k = (x_k, s_k, \kappa_k)$, $k \geq 0$ starting at $(x_0, s_0, 0)$.

Proof. The proof follows from the general stability result Theorem 5.1 stated in Section 5 and is postponed to Section 6. □

Remark 1. *Note that condition (1) is easily checked for $\Delta_k = O(k^{-\alpha})$ and $\epsilon_k = O(k^{-\alpha'})$ with $1/2 < \alpha' < \alpha < 1$. However condition (2) is somewhat less tractable and should be relaxed in future work.*

Remark 2. *Note that as observed in [5] (Lemma 2), since $\hat{\theta}$, ϕ and ψ are smooth, we get from* (3.5) *and* (3.7) *that if $\mathcal{L} \triangleq \{\ \theta \in \hat{\theta}(\mathcal{S}),\ \frac{\partial l}{\partial \theta}(\theta) = 0\}$, then $\hat{\theta}(\mathcal{L}') = \mathcal{L}$ and $\lim\limits_{k\to\infty} d(\theta_k, \mathcal{L}) = 0$ $\bar{\mathbb{P}}_{x_0,s_0}$-a.s*

4. Experiments.

To illustrate our stochastic algorithm for the deformable template models, we consider handwritten digit images. For each digit class, we learn the template, the corresponding noise variance and the geometric covariance matrices (note that in the experiments the noise variance is no longer fixed and is estimated as the other parameters). We use the US-Postal database which contains a training set of around 7000 images and a test set of 2007 images.

Each picture is a (16×16) gray level image with intensity in $[0, 2]$ where 0 corresponds to the black background. We will also use these sets in the special case of a noisy setting by adding independent normalised Gaussian noise to each image.

To be able to compare the results with the previous deterministic algorithm proposed in [1], we use the same samples. In Figure (4.1) below, we show some of the training images used for the statistical estimation.

A natural choice for the prior laws on α and Γ_g is to set 0 for the mean on α and to induce the two covariance matrices by the metric of the spaces V_p and V_g involving the correlation between

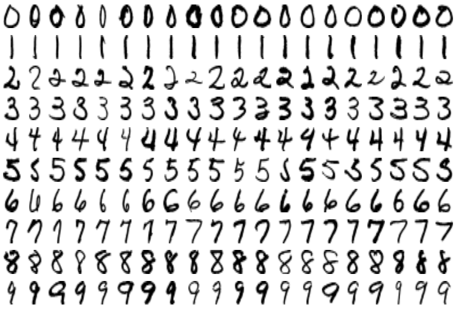

FIG. 4.1. *Training set used for the estimation of the model parameters.*

FIG. 4.2. *Estimated prototypes of digit 1 (20 images per class) for different hyper-parameters. Left: smoother geometry but large photometric covariance in the spline kernel. Right: more rigid geometry and smaller photometric covariance.*

the landmarks determined by the kernel. Define the square matrices

$$\begin{aligned} M_p(k,k') &= K_p(r_{p,k}, r_{p,k'}) \ \forall 1 \leq k,k' \leq k_p \\ M_g(k,k') &= K_g(r_{g,k}, r_{g,k'}) \ \forall 1 \leq k,k' \leq k_g \end{aligned} \tag{4.1}$$

then $\Sigma_p = M_p^{-1}$ and $\Sigma_g = M_g^{-1}$. In our experiments, we have chosen Gaussian kernels for both K_p and K_g, where the standard deviations are fixed at $\sigma_p = 0.12$ and $\sigma_g = 0.3$. These two variances are some important parameters; indeed, it has been shown in [1] that changing the geometrical covariance had an effect on the sharpness of the template images. Concerning the effect of the photometrical hyper-parameter, it affects both the template and the geometry in the sense that with a too large variance, the kernel centred on one landmark spreads out on too many of its neighbours. This leads to some thicker shapes as shown in left panel of Figure (4.2). As a consequence, the template is biased: it is not "centred" in the sense that the mean of the deformations required to fit the data is not close to zero. For example for the digit "1", the main deformations should be some contractions or dilations of the template. With a large variance σ_p^2, the template is thicker yielding larger contractions and smaller dilations. Since we have set a Gaussian law on the deformation variable β and the spline model of the deformation is anti-symmetric ($z_{-\beta} = -z_\beta$), for each deformation ($Id + z_\beta$) learnt, its symmetric deformation ($Id - z_\beta$) will be learnt as well. Looking at some synthetic examples given in Figure (4.3) top panel, there are many large dilated shapes whereas these examples were not in the training set and does not appear with the other hyper-parameters (Figure (4.3) bottom panel). This particular effect is due to the model we set for the template; indeed, the spline model requires some landmarks on the domain and the variance of the kernel K_p has to be fixed according to the distance between landmarks (and the kind of images treated). We have tried different relevant values and kept the best with regard to the visual results. We present in the following only the results with the adapted variances.

For the stochastic approximation step-size, we allow a heating period which corresponds to the absence of memory for the first iterations. This allows the Markov Chain to reach a region of interest in the posterior probability density function $q(\beta|y)$ before exploring this particular region.

FIG. 4.3. *Synthetic examples corresponding to the two previous templates of digit 1.*

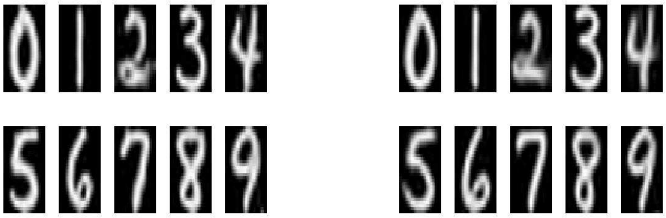

FIG. 4.4. *Estimated prototypes issued from left 10 images per class and right 20 images per class in the training set.*

In the experiments run here, the heating time lasts k_h (up to 150) iterations and the whole algorithm is stopped at, at most, 200 iterations depending on the data set (noisy or not). This number of iterations corresponds to a point where the convergence seems to be reached. This yields:

$$\Delta_k = \begin{cases} 1 \; \forall 1 \; \leq k \leq k_h \\ \frac{1}{(k-k_h)^d} \; \forall k > k_h \end{cases} \text{for } d = 0.6 \text{ or } 1 \; .$$

4.1. Estimated Template. We show here the results of the statistical learning algorithm for the one component model. Ten images per class are enough to obtain very contrasted and satisfactory template images. Increasing the number of training images does not significantly improve the estimated photometric template and may at some point provoke some deterioration of the templates. Indeed, if there are only few images, the model will fit these data precisely but as soon as some "outliers" appear the model will try to explain them as well by enlarging the estimated variability. The resulting estimated parameters can thus be less accurate. Figure (4.4) shows two runs of the one component algorithm for a non noisy data base with respectively 10 and 20 images per class.

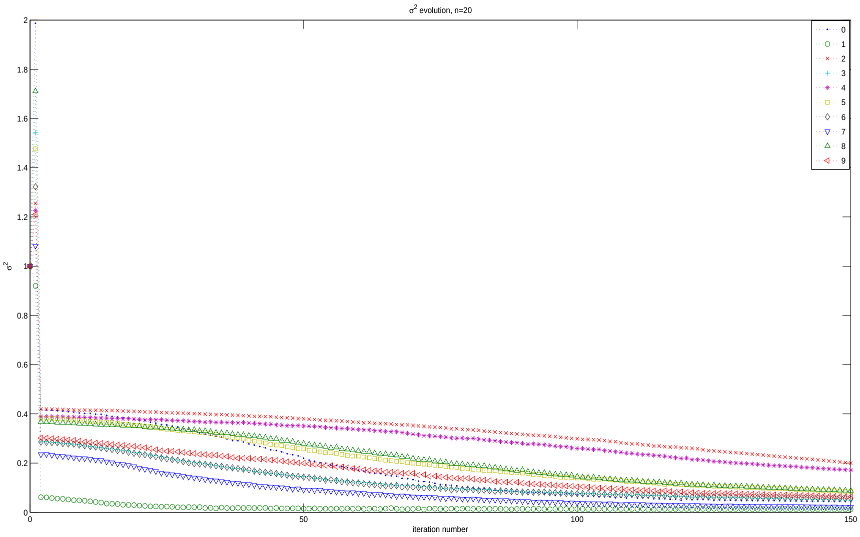

FIG. 4.5. *Estimated noise variance using* 20 *images per class.*

4.2. Photometric noise variance. The same behaviour for our stochastic EM as for the mode approximation EM algorithm done in [1] is observed for the noise variance: during the first iterations, the noise variance balances the inaccuracy of the estimated template which is simply the gray-level mean of the training set. As the iterations proceed, the templates estimates become more precise as does the estimate of covariance matrix for the geometry. This yields very small residual noise. Note that here the final noise variances are smaller than for the mode approximation; between 0.2 and 0.3 for the mode in the one component run and less than 0.1 for the Stochastic EM for all digits. This can be explained by the stochastic nature of the algorithm which enables it to escape from local minima provoking early stops in the deterministic version.

4.3. Estimated geometric distribution. As said previously we have to fix the value of the hyper-parameter a_g of the prior on Γ_g. This quantity has a significant role in the results. Indeed, to satisfy the theoretical conditions we have to choose a_g larger than $4k_g + 1$ that is to say $4 \times 36 + 1$ in our examples. But if we have a look at the geometry update equation which is a barycenter between what we have learnt and the prior with coefficients equal to the number n of images and a_g respectively, we notice that with a small number of images in the training set, the prior will dominate. This will not allow the covariance matrix to move away from that prior. We thus need to decrease a_g and find the best trade-off between the degenerate inverse Wishart and the weight of the prior in the covariance estimation. We fix this value with a visual criterion: both the templates and the generated sample with the learnt geometry have to be satisfactory. This yields $a_g = 0.5$ or 0.1.

We do note however that the fact that the prior is degenerated does not really matter as soon as the posterior distribution is not. In addition, considering the update formulas, even if this law does not have a total weight equal to 1 (for it to be a probability distribution) it does not affect the parameter estimation.

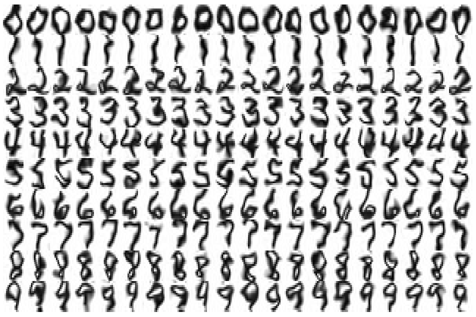

FIG. 4.6. *20 synthetic examples per class generated with the estimated template but the prior covariance matrix.*

In Figure (4.7), we show a sample of some synthetic digits drawn with respect to the model with the estimated parameters. Note that the resulting digits in Figure (4.7) look like some elements of the training set and seem to explain it correctly. In particular, for some especially geometrically constrained digits such as 0 or 1, the geometry variability reflects their constrains. For digits like the 2s, the training set is heterogeneous and shows a large geometrical variability. Comparing to the deformations obtained by the mode approximation in [1], it seems that here we obtain a less rigid geometry. This might be because with a stochastic algorithm, we explore the posterior density and do not only concentrate at its mode. This allows some more exotic deformations corresponding to realizations of the missing variable β which may belong to the tail of the law. Another reason may be that for such digits, the mode approximation gets stuck in a local minimum of the matching energy. Jumping out of this configuration would require a large deformation (not allowed by the gradient descent since it would increase the energy again) whereas such a deformation can be proposed and accepted by the stochastic algorithm. Subsequently the deformed template may better fit the observations leading to acceptance of these large deformations. This also leads to a lower value of the residual noise and may also explain the low noise variance estimated by the

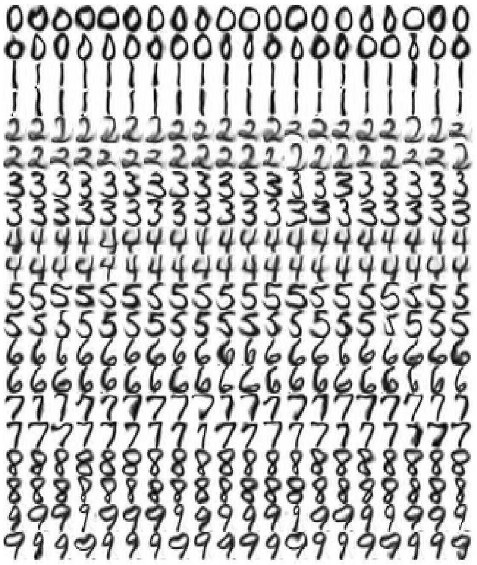

FIG. 4.7. *40 synthetic examples per class generated with the estimated parameters: 20 with the direct deformations and 20 with the symmetric deformations.*

stochastic EM algorithm.

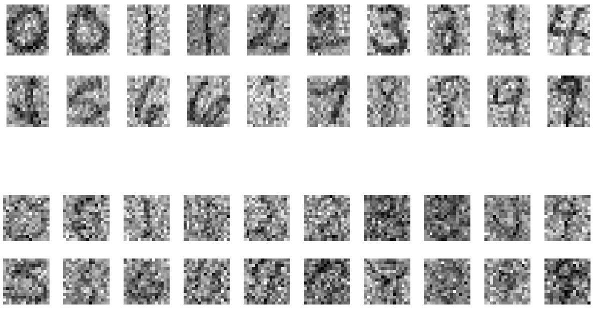

FIG. 4.8. *Two images examples per class of the noisy training set (variance: top: $\sigma^2 = 1$, bottom: $\sigma^2 = 2$).*

4.4. Noise effect. As shown in [1], in the presence of noise, the mode approximation algorithm does not converge toward the MAP estimator. In our setting, the consistence of the "SAEM like" algorithm has been proved independently of the training set, thus noisy images can also be treated exactly the same way. These are the results we present here. Figure (4.8) shows two training examples per class for noise variance values $\sigma^2 = 1$ and $\sigma^2 = 2$. In Figures (4.9) and (4.10), we show the estimated templates for the noisy training set containing 20 images for both methods. Even if the mode approximation algorithm does not diverge, it cannot fit the template for digits with a high variability whereas the stochastic EM finds the template and gives acceptable contrasted templates which look like those obtained in Figure (4.4). This becomes more significant as we increase the variance of the additive noise we introduce in the training set.

The same choice of the hyper-parameters has to be done. For the geometry, there is not reason to change them. Concerning the photometric variance of the spline kernel, a too small one could

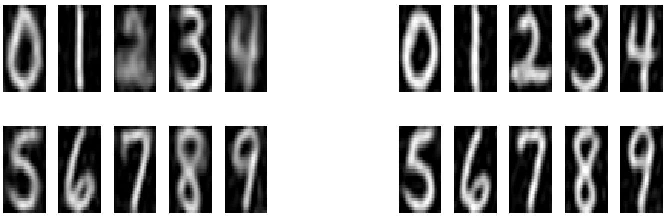

FIG. 4.9. *Estimated prototypes in a noisy setting $\sigma^2 = 1$: Left: with the mode approximation algorithm. Right: with the SAEM-MCMC coupling procedure.*

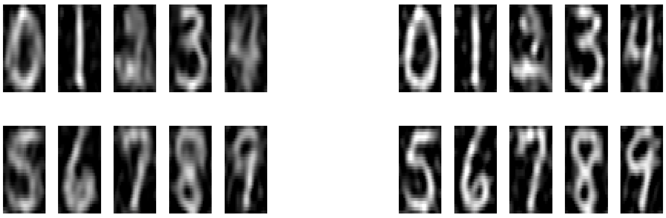

FIG. 4.10. *Estimated prototypes in a noisy setting $\sigma^2 = 2$: Left: with the mode approximation algorithm. Right: with the SAEM-MCMC coupling procedure.*

create some non smooth templates whereas a larger kernel would smooth the noise effect. We are presenting here only some experiments which seem to be a good tradeoff between these effects.

The geometry is also well estimated despite the high level of noise in the training set. Figure (4.11) shows some synthetic examples drawn with the estimated parameters learnt from the noisy training set with an additive noise variance of 1. The two lines correspond to deformations and their symmetric deformation. This sample looks like the synthetic samples learnt on non noisy images even if some example are not really relevant. However, the global behaviour has been learnt.

The algorithm manages to catch the photometry (a contrasted and smooth template) and the geometry of the shapes and to "separate" the additive noise.

The number of iterations needed to reach the convergence point in the noisy setting is about twice that of the non noisy case. The convergence of the template is the longest whereas the convergence of σ^2 takes relatively the same number of iterations. In particular, the templates obtained in the left panel of Figure (4.4) with only 10 images per training digit set are obtained with a heating period of 25 iterations and 5 more EM steps with memory. Whereas the templates of Figure (4.9) left picture require 100 to 125 heating iterations in the 150 EM iterations. This is understandable since the algorithm has to cope with variations due to the noise and thus needs a longer time to fit the right model.

5. Main stochastic approximation convergence Theorem. We give here a theorem that, under some assumptions, will ensure the convergence of the stochastic approximation sequence $(s_k)_k$. This is a direct adaptation of the convergence theorem of [3] with weaker Hölder conditions.

We consider the following assumptions, generalised from [3]. Define for $V : \mathbb{R}^N \to [1, \infty[$ and any $g : \mathbb{R}^N \to \mathbb{R}^{n_s}$ the norm

$$\|g\|_V = \sup_{x \in \mathbb{R}^N} \frac{|g(x)|}{V(x)} \tag{5.1}$$

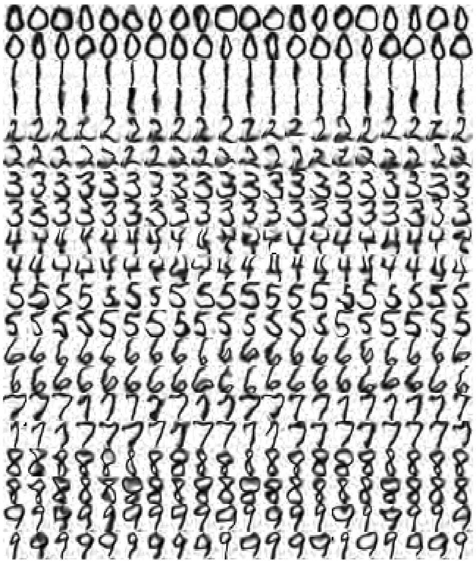

FIG. 4.11. *40 synthetic examples per class generated with the parameters estimated from the noisy training set (additive noise variance of* 1*).*

A1' $\mathcal{S}$ is an open subset of $\mathbb{R}^{n_s}$, $h : \mathcal{S} \to \mathbb{R}^{n_s}$ is continuous and there exists a continuously differentiable function $w : \mathcal{S} \to [0, \infty[$ such that

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

$$\mathcal{L}' \triangleq \{s \in \mathcal{S}, \langle \nabla w(s), h(s) \rangle = 0\} \subset \{s \in \mathcal{S},\ w(s) < M_0\},$$

(ii) There exist a closed convex set $\mathcal{S}_a \subset \mathcal{S}$ for which $s \to \rho H(s, x) \in \mathcal{S}_a$ for any $\rho \in [0, 1]$ and $(s, x) \in \mathcal{S}_a \times \mathbb{R}^N$ ($\mathcal{S}_a$ is absorbing) such that for any $M_1 \in]M_0, \infty]$, we have $\mathcal{W}_{M_1} \cap \mathcal{S}_a$ is a compact set of $\mathcal{S}$ where $\mathcal{W}_{M_1} \triangleq \{s \in \mathcal{S},\ w(s) \leq M_1\}$,

(iii) For any $s \in \mathcal{S} \backslash \mathcal{L}'$ $\langle \nabla w(s), h(s) \rangle < 0$,

(iv) The closure of $w(\mathcal{L}')$ has an empty interior.

A2 For any $\theta \in \hat{\theta}(\mathcal{S})$, the Markov kernel Π_θ has a single stationary distribution π_θ, $\pi_\theta \Pi_\theta = \pi_\theta$. In addition $H : \mathcal{S} \times \mathbb{R}^N \to \mathcal{S}$ is measurable, for all $s \in \mathcal{S}$, $\int_{\mathbb{R}^N} |H(s, x)| \pi_{\hat{\theta}(s)}(dx) < \infty$.

A3' For any $s \in \mathcal{S}$ and $\theta = \hat{\theta}(s)$, the Poisson equation $g - \Pi_\theta g = H_s - \pi_\theta(H_s)$ where $H_s(x) \triangleq H(s, x)$ has a solution g_s. There exists a function $V : \mathbb{R}^N \to [1, \infty]$ such that $\{x \in \mathbb{R}^N, V(x) < \infty\} \neq \emptyset$, a constant $a \in]0, 1]$ and an integer $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \mathcal{S}$,

(i)

$$\sup_{s \in \mathcal{K}} \|H_s\|_V < \infty \tag{5.2}$$

$$\sup_{s \in \mathcal{K}} (\|g_s\|_V + \|\Pi_{\hat{\theta}(s)} g_s\|_V) < \infty \tag{5.3}$$

(ii)

$$\sup_{s, s' \in \mathcal{K}} |s - s'|^{-a} \{\|g_s - g_{s'}\|_{V^{3/2}} + \|\Pi_{\hat{\theta}(s)} g_s - \Pi_{\hat{\theta}(s')} g_{s'}\|_{V^{3/2}}) < \infty \tag{5.4}$$

(iii) Let k_0 be an integer; there exist $\epsilon > 0$ and a constant C such that for any sequence $(\varepsilon_k)_k$ satisfying $0 < \varepsilon_k \leq \varepsilon$ for all $k \geq k_0$, for any sequence $\Delta = (\Delta_k)_{k\geq 0}$ and for any $x \in \mathbb{R}^N$,

$$\sup_{s\in\mathcal{K}} \sup_{k\geq 0} \mathbb{E}^{\Delta}_{x,s} \left[V^p(x_k) \mathbb{1}_{\sigma(\mathcal{K})\wedge\nu_\varepsilon \geq k} \right] \leq C V^p(x) \tag{5.5}$$

where $\nu_\varepsilon = \inf\{k \geq 1, |s_k - s_{k-1}| \geq \varepsilon_k\}$ and $\sigma(\mathcal{K}) = \inf\{k \geq 1, s_k \notin \mathcal{K}\}$ and the expectation is related to the non-homogeneous Markov Chain $(x_k, s_k)_{k\geq 0}$ with step-size sequence $(\Delta_k)_{k\geq 0}$.

A4 The sequences $\Delta = (\Delta_k)_{k\geq 0}$ and $\varepsilon = (\varepsilon_k)_{k\geq 0}$ are non-increasing, positive and satisfy: $\sum\limits_{k=0}^{\infty} \Delta_k = \infty$, $\lim\limits_{k\to\infty} \varepsilon_k = 0$ and $\sum\limits_{k=1}^{\infty} \{\Delta_k^2 + \Delta_k \varepsilon_k^a + (\Delta_k \varepsilon_k^{-1})^p\} < \infty$ where a and p are defined in (**A3'**).

Theorem 5.1 (General Convergence Result for Truncated Stochastic Approximation). *Assume (**A1'**),(**A2**), (**A3'**) and (**A4**). Let* $\mathrm{K} \subset \mathbb{R}^N$ *such that* $\sup\limits_{x\in\mathrm{K}} V(x) < \infty$ *and* $\mathcal{K}_0 \subset \mathcal{W}_{M_0} \cap \mathcal{S}_a$ *(where* M_0 *is defined in (**A1'**)), and let* $(s_k)_{k\geq 0}$ *be the sequence defined in Algorithm 1. Then, for all* $x_0 \in \mathrm{K}$ *and* $s_0 \in \mathcal{K}_0$*, we have* $\lim\limits_{k\to\infty} d(s_k, \mathcal{L}') = 0$ $\bar{\mathbb{P}}_{x_0,s_0}$*-a.s, where* $\bar{\mathbb{P}}_{x_0,s_0}$ *is the probability measure associated with the chain* $Z_k = (x_k, s_k, \kappa_k)$, $k \geq 0$ *starting at* $(x_0, s_0, 0)$.

The proof that we satisfy assumptions (**A1'**),(**A2**), (**A3'**) and (**A4**) is given in Section 6. The convergence of the sequence $(s_k)_k$ is a consequence of Theorem 5.5 of [3] with these assumptions.

Proof. • The deterministic results obtained by [3] under their assumption (**A1**) remain true if we suppose the existence of an absorbing set as defined in assumption (**A1'**). Indeed, the proofs can be carried through in the same way restricting the sequences to the absorbing set.

- Assumption (**A2**) remains unchanged.
- We have to prove an equivalent of proposition 5.2 from [3].

Proposition 5.2. *Assume (**A3'**). Let* $\mathcal{K}$ *be a compact subset of* $\mathcal{S}$ *and let* $\Delta = (\Delta_k)_k$ *and* $\varepsilon = (\varepsilon_k)_k$ *be two non-increasing sequences of positive numbers such that* $\lim\limits_{k\to\infty} \varepsilon_k = 0$*. Then, for p as defined in (**A3'**),*

1 *There exists a constant* C *such that, for any* $(x, s) \in \mathbb{R}^N \times \mathcal{K}$ *and any integer* l*, any* $\delta > 0$

$$\mathbb{P}^{\Delta}_{x,s}\left(\sup_{n\geq l} |S_{l,n}(\varepsilon, \Delta, \mathcal{K})| \geq \delta\right) \leq C\delta^{-p} \left\{ \left(\sum_{k=l}^{\infty} \Delta_k^2\right)^{p/2} + \left(\sum_{k=l}^{\infty} \Delta_k \varepsilon_k^a\right)^p \right\} V^{3p/2}(x)$$

where $S_{l,n}(\varepsilon, \Delta, \mathcal{K}) \triangleq \mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\varepsilon)\geq n} \sum\limits_{k=l}^{n} \Delta_k (H(s_{k-1}, x_k) - h(s_{k-1}))$ *and* $\mathbb{P}^{\Delta}_{x,s}$ *is the probability measure generated by the non homogeneous Markov Chain* $((x_k, s_k))_k$ *started from the initial condition* (x, s).

2 *There exists a constant* C *such that for any* $(x, s) \in \mathbb{R}^N \times \mathcal{K}$

$$\mathbb{P}^{\Delta}_{x,s}(\nu(\varepsilon) < \sigma(\mathcal{K})) \leq C \left\{ \sum_{k=l}^{\infty} (\Delta_k \varepsilon_k^{-1})^p \right\} V^{3p/2}(x) \ . \tag{5.6}$$

The proof of this proposition can proceed as in [3] except for the upper bound of the term involving the Hölder property. Under (**A3'(ii)**), this upper bound brings into play an exponent $3p/2$ on the function V. This leads to the two previous majorations in the previous proposition.

Given this proposition, it is straightforward to prove the following proposition which corresponds to Proposition 5.3 in [3].

Proposition 5.3. *Assume (**A3'**) and (**A4**). Then, for any subset* $\mathrm{K} \subset \mathbb{R}^N$ *such that* $\sup\limits_{x\in\mathrm{K}} V(x) < \infty$*, any* $M \in (M_0, M_1]$ *and any* $\delta > 0$*, we have* $\lim\limits_{k\to\infty} A(\delta, \varepsilon^{\leftarrow k}, M, \Delta^{\leftarrow k}) = 0$ *where*

$$A(\delta, \varepsilon, M, \Delta) \quad = \quad \sup_{s\in\mathcal{K}_0} \sup_{x\in\mathrm{K}} \left\{ \mathbb{P}^{\Delta}_{x,s}\left(\sup_{k\geq 1} |S_{1,k}(\varepsilon, \Delta, \mathcal{W}_M)| \geq \delta\right) + \mathbb{P}^{\Delta}_{x,s}\left(\nu(\varepsilon) < \sigma(\mathcal{W}_M)\right) \right\} \quad .$$

The convergence of the sequence $(s_k)_k$ follows from the proof of Theorem 5.5 of [3] with our assumptions.

□

6. Proof of the convergence of the truncated SAEM/MCMC algorithm for deformable template. Here we demonstrate Theorem 3.1 that is to say the convergence of the parameter sequence obtained by the coupling procedure for one component model. We recall that in this Section, the parameter σ^2 is fixed so that $\theta = (\alpha, \Gamma)$. The sufficient statistic vector S, the set $\mathcal{S}$ as well as the explicit expression of $\hat{\theta}(s)$ have been given in Subsection 3.4. As noted, $\hat{\theta}$ is a smooth function of $\mathcal{S}$.

We will prove the conditions (**A1'**), (**A2**), (**A3'**) and (**A4**) hold for any $p \geq 1$ and $a \in]0,1[$ so that applying Theorem 5.1, we get immediately Theorem 3.1.

6.1. (A1'). We choose the same functions H, h and w as in [5] defined as follows:

$$\begin{aligned} H(s,x) &= H_s(x) = S(x) - s \\ h(s) &= \int_{\mathbb{R}^N} H(s,x) q(x|y,\theta) dx \\ w(s) &= -l(\hat{\theta}(s)) \ . \end{aligned}$$

where as introduced before, x stands for the family β_1^n and y for y_1^n. As shown in [5], we get (**A1'(iii)**) and (**A1'(iv)**).

Moreover, there exists an absorbing closed subset $\mathcal{S}_a$ of $\mathbb{R}^{n_s}$ such that

$$s + \rho H(s,x) \in \mathcal{S}_a \ \text{ for any } \rho \in [0,1] \text{ and } s \in \mathcal{S}_a. \tag{6.1}$$

Indeed, since the interpolation kernel K_p is bounded, there exist $a > 0$ and $A \in \mathrm{Sym}^+_{k_p}$ such that for any $x \in \mathbb{R}^N$, we have

$$|S_1(x)| \leq a, \ 0 \leq S_2(x) \leq A \text{ and } 0 \leq S_3(x) \tag{6.2}$$

where, as usual, for any symmetric matrices B and C, we say that $B \leq C$ if $C - B$ is a non-negative symmetric matrix.

Now define the set $\mathcal{S}_a$ by

$$\mathcal{S}_a \triangleq \{ \ s \in \mathcal{S} \ | \ |s_1| \leq a, \ 0 \leq s_2 \leq A \ \text{ and } 0 \leq s_3 \ \} \, .$$

Since the constraints are obviously convex and closed, we get that $\mathcal{S}_a$ is a closed convex subset of $\mathbb{R}^{n_s}$ such that

$$\mathcal{S}_a \subset \mathcal{S} \subset \mathbb{R}^{n_s}$$

and satisfying (6.1).

We now focus on the first two points. As l and $\hat{\theta}$ are continuous functions, we only need to prove that $\mathcal{W}_M \cap \mathcal{S}_a$ is a bounded set for a constant $M \in \mathbb{R}_+^*$ with:

$$\mathcal{W}_M = \{ s \in \mathcal{S}, \ -l(\hat{\theta}(s)) \leq M \} \, .$$

On $\mathcal{S}_a$, s_1 and s_2 are bounded so that if $\hat{\theta}(s) = (\alpha(s), \Gamma(s))$, we deduce from (3.14) and from the boundedness of interpolation kernel K_p that $\alpha(s)$ is bounded on $\mathcal{S}_a$ and $|y_i - K_p^{\beta_i}\alpha(s)|$ is uniformly bounded on $\beta \in \mathbb{R}^{2k_g}$ and $s \in \mathcal{S}_a$. Hence (recall that σ^2 is fixed here), there exists $\eta > 0$ such that $q(y|x, \hat{\theta}(s)) \geq \eta$ for any $s \in \mathcal{S}_a$ and $x \in \mathbb{R}^N$. We have

$$w(s) \geq -\log(\int q(x, \hat{\theta}(s)) dx) + \mathrm{C} \geq -\log(q(\hat{\theta}(s))) + \mathrm{C} \geq -\log(q(\Gamma(s))) + \mathrm{C}' \tag{6.3}$$

where C and C′ are two constants independent of $s \in S_a$. Since

$$-\log(q(\Gamma_g)) = \frac{a_g}{2}\left(\langle \Gamma_g^{-1}, \Sigma_g\rangle_F + \log|\Gamma_g|\right) \geq \frac{a_g}{2}\log|\Gamma_g|$$

and $\log(|\Gamma_g(s)|) = \log(|(s_3 + a_g\Sigma_g)/(n+a_g)|) \to +\infty$ as $|s| \to +\infty$, $s \in \mathcal{S}_a$, we deduce that

$$\lim_{|s|\to+\infty, s\in\mathcal{S}_a} w(s) = +\infty\,.$$

Since w is continuous and $\mathcal{S}_a$ is closed, this proves (**A1'(ii)**).

Considering (**A1'(i)**), we assume that the assumption is satisfied.

6.2. (A2). We prove a classical sufficient condition (**DRI1**), used in [3] which will imply (**A2**).

(DRI1) For any $s \in \mathcal{S}$, $\Pi_{\hat\theta(s)}$ is irreducible and aperiodic. In addition there exist a small set $\mathtt{C}$ (defined below), a function $V : \mathbb{R}^N \to [1, \infty[$ and constants $0 \leq b \leq 1$, such that , for any $p \geq 2$ and any compact subset $\mathcal{K} \subset \mathcal{S}$, there exist an integer m and constants $0 < \lambda < 1$, B, κ, $\delta > 0$ and a probability measure ν such that

$$\sup_{s\in\mathcal{K}} \Pi^m_{\hat\theta(s)} V^p(x) \leq \lambda V^p(x) + B\mathbb{1}_{\mathtt{C}}(x) \tag{6.4}$$

$$\sup_{s\in\mathcal{K}} \Pi_{\hat\theta(s)} V^p(x) \leq \kappa V^p(x) \quad \forall x \in \mathbb{R}^N \tag{6.5}$$

$$\sup_{s\in\mathcal{K}} \Pi^m_{\hat\theta(s)}(x, A) \geq \delta\nu(A) \quad \forall x \in \mathtt{C}, \forall A \in \mathcal{B}(\mathbb{R}^N)\ . \tag{6.6}$$

NOTATION 2. *Let $(e_j)_{1\leq j\leq N}$ be the canonical basis of the x-space and for any $1 \leq j \leq N$, let $E_{\theta,j} \triangleq \{\ x \in \mathbb{R}^N \mid \langle x, e_j\rangle_\theta = 0\}$ be the orthogonal of Span$\{e_j\}$ and $p_{\theta,j}$ be the orthogonal projection on $E_{\theta,j}$ i.e.*

$$p_{\theta,j}(x) \triangleq x - \frac{\langle x, e_j\rangle_\theta}{|e_j|^2_\theta} e_j\,,$$

where $\langle x, x'\rangle_\theta = \sum_{i=1}^n \beta_i^t \Gamma_g^{-1}\beta_i$ for $\theta = (\alpha, \Gamma_g)$ and $x = \beta_1^n$, $x' = \beta'^n_1$ (i.e. the natural dot product associated with the covariance matrix Γ_g).

We denote for any $1 \leq j \leq N$ and $\theta \in \Theta$, $\Pi_{\theta,j}$ the Markov kernel on $\mathbb{R}^N$ associated with the j-th Hasting-Metropolis step of the Gibbs sampler on x. We have $\Pi_\theta = \Pi_{\theta,N} \circ \cdots \Pi_{\theta,1}$.

We first recall the definition of a small set:

DEFINITION 1. *([11]) A set $\mathcal{E} \in \mathcal{B}(\mathcal{X})$ is called a* ***small set*** *for the kernel Π if there exists an $m > 0$, and a non trivial measure ν_m on $\mathcal{B}(\mathcal{X})$, such that for all $x \in \mathcal{E}$, $B \in \mathcal{B}(\mathcal{X})$,*

$$\Pi^m(x, B) \geq \nu_m(B). \tag{6.7}$$

When (6.7) holds, we say that $\mathcal{E}$ is ν_m-small. We now prove the following lemma:

LEMMA 6.1. *Let $\mathcal{E}$ be a compact subset of $\mathbb{R}^N$ then $\mathcal{E}$ is a small set of $\mathbb{R}^N$ for $(\Pi_{\hat\theta(s)})_{s\in\mathcal{K}}$.*

Proof. First note that there exists $a_c > 0$ such that for any $\theta \in \Theta$, any $x \in \mathbb{R}^N$ and any $b \in \mathbb{R}$, the acceptance rate $r_{\theta,j}(x, b)$ is uniformly lower bounded by a_c so that for any $1 \leq j \leq N$ and any non-negative function f,

$$\Pi_{\theta,j} f(x) \geq a_c \int_{\mathbb{R}} f(x_{-j} + be_j) q(b|x_{-j}, \theta) db = a_c \int_{\mathbb{R}} f(p_{\theta,j}(x) + ze_j/|e_j|_\theta) g_{0,1}(z) dz$$

where $g_{0,1}$ is the standard $\mathcal{N}(0, 1)$ density.

By induction, we have

$$\Pi_\theta f(x) \geq a_c^N \int_{\mathbb{R}^N} f\left(p_{\theta,N,1}(x) + \sum_{j=1}^N z_j p_{\theta,N,j+1}(e_j)/|e_j|_\theta\right) \prod_{j=1}^N g_{0,1}(z_j) dz_j \tag{6.8}$$

where $p_{\theta,q,r} = p_{\theta,r} \circ p_{\theta,r-1} \circ \cdots \circ p_{\theta,q}$ for any integer $q \leq r$ $p_{\theta,N,N+1} = \mathrm{Id}_{\mathbb{R}^N}$. Let $A_\theta \in \mathcal{L}(\mathbb{R}^N)$ be the linear mapping on $z_1^N = (z_1, \cdots, z_N)$ defined by $A_\theta z_1^N = \sum_{j=1}^N z_j p_{\theta,N,j+1}(e_j)/|e_j|_\theta$. One easily checks that for any $1 \leq k \leq N$, $\mathrm{Span}\{\ p_{\theta,N,j+1}(e_j),\ k \leq j \leq N\} = \mathrm{Span}\{e_j \mid k \leq j \leq N\}$ so that A_θ is an invertible mapping. By a change of variable, we get

$$\int_{\mathbb{R}^N} f\left(p_{\theta,N,1}(x) + A_\theta z_1^N\right) \prod_{j=1}^N g_{0,1}(z_j) dz_j = \int_{\mathbb{R}^N} f(u) g_{p_{\theta,N,1}(x), A_\theta A_\theta^t}(u) du$$

where $g_{\mu,\Sigma}$ stands for the density of the normal law $\mathcal{N}(\mu, \Sigma)$. Since $\theta \to A_\theta$ is smooth on the set of invertible mappings in θ, we deduce that there exists $c > 0$ such that $c\mathrm{Id} \leq A_\theta A_\theta^t \leq \mathrm{Id}/c$ and $g_{p_{\theta,N,1}(x), A_\theta A_\theta^t}(u) \geq C g_{p_{\theta,N,1}(x), \mathrm{Id}/c}(u)$ uniformly for $\theta = \hat{\theta}(s)$ with $s \in \mathcal{K}$. Assuming that $x \in \mathcal{E}$, since $\theta \to p_{\theta,N,1}$ is smooth and $\mathcal{E}$ is compact, we have $\sup_{x\in\mathcal{E}, \theta=\hat{\theta}(s),\ s\in\mathcal{K}} |p_{\theta,N,1}(x)| < \infty$ so that there exists $C' > 0$ and $c' > 0$ such that for any $(u, x) \in \mathbb{R}^N \times \mathcal{E}$ and any $\theta = \hat{\theta}(s),\ s \in \mathcal{K}$

$$g_{p_{\theta,N,1}(x), A_\theta A_\theta^t}(u) \geq C' g_{0,\mathrm{Id}/c'}(u). \tag{6.9}$$

Using (6.8) and (6.9), we deduce that for any A

$$\Pi_\theta(x, A) \geq C' a_c^N \nu(A)$$

with ν equal to the density of the normal law $\mathcal{N}(0, \mathrm{Id}/c')$.

This yields the existence of the small set as well as equation (6.6). □

This property also implies the ϕ-irreducibility of the Markov chain $(x_k)_k$.

Moreover, the existence of a ν_1-small set implies the aperiodicity of the chain (cf:[11] p121).

We set $V : \mathbb{R}^N \to [1, +\infty[$ as the following function

$$V(x) = 1 + \sum_{i=1}^n |\beta_i|^2. \tag{6.10}$$

We now prove condition (6.5). For any $1 \leq j \leq N$ and any θ, we have

$$\Pi_{\theta,j} V^p(x) \leq V^p(x) + \int_{\mathbb{R}} V^p(p_{\theta,j}(x) + z e_j/|e_j|_\theta) g_{0,1}(z) dz\,.$$

Since $V(x + h) \leq 2(V(x) + V(h))$ for any $x, h \in \mathbb{R}^N$, $|p_{\theta,j}(x)| \leq C|x|$ and $|e_j|_\theta \geq 1/c$ for C and $c > 0$ independent of $\theta = \theta(s), s \in \mathcal{K}$, we have

$$\int_{\mathbb{R}} V^p(p_{\theta,j}(x) + z e_j/|e_j|_\theta) g_{0,1}(z) dz \leq 2^p C^p V^p(|x|) \int_{\mathbb{R}} (1 + V(cze_j))^p g_{0,1}(z) dz$$

we deduce that there exists $C' > 0$ such that for any $x \in \mathbb{R}^N$

$$\sup_{\theta=\theta(s), s\in\mathcal{K}} \Pi_{\theta,j} V^p(x) \leq C' V^p(x)\,, \tag{6.11}$$

such that by composition $\Pi_\theta V^p(x) \leq C'^N V^p(x)$ and (6.5) holds.

Now consider the Drift condition (6.4).

For any $\theta = (\alpha, \Gamma_g)$, we introduce a θ dependent function $V_\theta(x) \triangleq 1 + \sum_{i=1}^n |\beta_i|_\theta^2$, where $|\beta|_\theta^2 \triangleq \langle \beta, \beta \rangle_\theta = \beta^t \Gamma_g^{-1} \beta$ is the natural dot product induced by the covariance operator Γ_g.

Lemma 6.2. *Let K be a compact subset of Θ. For any integer $p \geq 1$, there exist $0 \leq \rho < 1$ and $C > 0$ such that for any $\theta \in K$, any $x \in \mathbb{R}^N$ we have*

$$\Pi_\theta V_\theta^p(x) \leq \rho V_\theta^p(x) + C\,.$$

Proof. The proposal distribution for $\Pi_{\theta,j}$ is given by $q(x \mid x_{-j}, y, \theta) \overset{\text{law}}{=} p_{\theta,j}(x) + U_\theta e_j$ where $U_\theta \sim \mathcal{N}(0, |e_j|_\theta^{-2})$. Since we easily check that the acceptance rate $a_{\theta,x}$ is uniformly bounded from below by a positive number $a_c > 0$, we deduce that there exists C_K such that for any $x \in \mathbb{R}^N$ and any measurable set $A \in \mathcal{B}(\mathbb{R}^N)$

$$\Pi_{\theta,j}(x, A) = (1 - a_{\theta,x})\mathbb{1}_A(x) + a_{\theta,x} \int_{\mathbb{R}} \mathbb{1}_A(p_{\theta,j}(x) + ze_j)\gamma_\theta(dz)$$

where $a_{\theta,x} \geq a_c$, $\gamma_\theta \leq C_K \gamma_K$ and γ_K equals to the density of the normal law $\mathcal{N}(0, \sup_{\theta \in K} |e_j|_\theta^{-2})$.

Since $\langle p_{\theta,j}(x), e_j \rangle_\theta = 0$, we get $V_\theta^p(p_{\theta,j}(x) + ze_j) = (V_\theta(p_{\theta,j}(x)) + z^2|e_j|_\theta^2)^p$ and

$$\begin{aligned}
\Pi_{\theta,j} V_\theta^p(x) &= (1 - a_{\theta,x})V_\theta^p(x) + a_{\theta,x} \int_{\mathbb{R}} \left(V_\theta(p_{\theta,j}(x)) + z^2|e_j|^2\right)^p \gamma_{\theta,x}(dz) \\
&\leq (1 - a_{\theta,x})V_\theta^p(x) + a_{\theta,x} \left(V_\theta^p(p_{\theta,j}(x)) + (2^p - 1)C_K V_\theta^{p-1}(p_{\theta,j}(x)) \int_{\mathbb{R}} (1 + z^2|e_j|_\theta^2)^{p-1} \gamma_K(dz) \right) \\
&\leq (1 - a_{\theta,x})V_\theta^p(x) + a_{\theta,x} V_\theta^p(p_{\theta,j}(x)) + C'_K V_\theta^{p-1}(p_{\theta,j}(x))
\end{aligned}$$

where we have used the fact that a Gaussian variable has bounded moment of any order. Since $a_{\theta,x} \geq a_c$ and $|p_{\theta,j}(x)|_\theta \leq |x|_\theta$ ($p_{\theta,j}$ is an orthonormal projection for the dot product $\langle \cdot, \cdot \rangle_\theta$), we get that for any $\eta > 0$, there exists $C_{K,\eta}$ such that for any $x \in \mathbb{R}^N$ and $\theta \in K$

$$\Pi_{\theta,j} V_\theta^p(x) \leq (1 - a_c)V_\theta^p(x) + (a_c + \eta)V_\theta^p(p_{\theta,j}(x)) + C_{K,\eta} \,.$$

By induction, starting with j=1, we get

$$\Pi_\theta V_\theta^p(x) \leq \sum_{u \in \{0,1\}^N} \prod_{j=1}^N (1 - a_c)^{1-u_j}(a_c + \eta)^{u_j} V_\theta^p(p_{\theta,u}(x)) + \frac{C_{K,\eta}}{\eta}((1 + \eta)^{N+1} - 1)$$

where $p_{\theta,u} = ((1 - u_N)\mathrm{Id} + u_N p_{\theta,N}) \circ \cdots \circ ((1 - u_1)\mathrm{Id} + u_1 p_{\theta,1})$. Let $p_\theta = p_{\theta,\mathbf{1}} = p_{\theta,N} \circ \cdots \circ p_{\theta,1}$ and note that $p_{\theta,u}$ is contracting so that

$$\Pi_\theta V_\theta^p(x) \leq b_{c,\eta} V_\theta^p(x) + (a_c + \eta)^N V_\theta^p(p_\theta(x)) + \frac{C_{K,\eta}}{\eta}((1 + \eta)^{N+1})$$

for $b_{c,\eta} = \left(\sum_{u \in \{0,1\}^N,\ u \neq \mathbf{1}} \prod_{j=1}^N (1 - a_c)^{1-u_j}(a_c + \eta)^{u_j} \right)$. To end the proof, we need to check that p_θ is strictly contracting uniformly on K. Indeed, $|p_\theta(x)|_\theta = |x|_\theta$ implies that $p_{\theta,j}(x) = x$ for any $1 \leq j \leq N$ so that $\langle x, e_j \rangle_\theta = 0$ and $x = 0$ since $(e_j)_{1 \leq j \leq N}$ is a basis. Using the continuity of the norm of p_θ in θ and the compactness of K, we deduce that there exists $0 < \rho_K < 1$ such that $|p_\theta(x)|_\theta \leq \rho_K |x|_\theta$ for any x and $\theta \in K$. Changing ρ_K for $1 > \rho'_K > \rho_K$ we get $(1 + \rho_K^2|x|_\theta^2)^q \leq \rho'^{2q}_K(1 + |x|_\theta^2)^q + C''_K$ for some uniform constant C''_K so that

$$\Pi_\theta V_\theta^p(x) \leq b_{c,\eta} V_\theta^p(x) + \rho'^{2p}_K (a_c + \eta)^N V_\theta^p(x) + C''_{K,\eta}.$$

Since we have $\inf_{\eta > 0} b_{c,\eta} + \rho'^{2p}_K(a_c + \eta)^N < 1$ we get the result. □

Lemma 6.3. *For any compact set $K \subset \Theta$, any integer $p \geq 0$, there exist $0 < \rho < 1$, $C > 0$ and m_0 such that $\forall m \geq m_0$, $\forall \theta \in K$*

$$\Pi_\theta^m V^p(x) \leq \rho V^p(x) + C \,.$$

Proof. Indeed, there exist $0 \leq c_1 \leq c_2$ such that $c_1 V(x) \leq V_\theta(x) \leq c_2 V(x)$ for any $(x, \theta) \in \mathbb{R}^N \times K$. Then, using the previous lemma, we have $P_\theta^m V^p(x) \leq c_1^{-p} P_\theta^m V_\theta^p(x) \leq c_1^{-p}(\rho^m V_\theta^p(x) + C/(1 - \rho)) \leq (c_2/c_1)^p(\rho^m V^p(x) + C/(1 - \rho))$. Choosing m large enough for $(c_2/c_1)^p \rho^m < 1$ gives the result. □

This finishes the proof of (6.4) and in the same time (**A2**).

6.3. (A3'). The geometric ergodicity of the Markov Chain, implied by the drift condition (6.4), ensures the existence of a solution of the Poisson equation (cf:[11]):

$$g_s(x) = \sum_{k\geq 0} (\Pi^k_{\hat{\theta}(s)} H_s(x) - h(s)). \tag{6.12}$$

We first focus on the proof of (**A3'(iii)**).

LEMMA 6.4. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. There exists $C > 0$ such that for any $s, s' \in \mathcal{K}$,*

$$|V^p_{\hat{\theta}(s)}(x) - V^p_{\hat{\theta}(s')}(x)| \leq C|s - s'| V^p_{\hat{\theta}(s)}(x)\,.$$

Proof. Indeed, there exists $C > 0$ such that for any $\hat{\theta}(s) = (\alpha, \Gamma_g)$ and $\hat{\theta}(s') = (\alpha', \Gamma'_g)$, $|\Gamma_g - \Gamma'_g| \leq C|s-s'|$. Therefore, there exists C' such that for any $s, s' \in \mathcal{K}$, $|\Gamma_g^{-1} - (\Gamma'_g)^{-1}| \leq C|s-s'|$ and $|V_{\hat{\theta}(s)}(x) - V_{\hat{\theta}(s')}(x)| \leq \sum_{i=1}^n \beta_i^t(\Gamma_g^{-1} - (\Gamma'_g)^{-1})\beta_i \leq C'|s-s'|V(x)$. Since there exists C'' such that $V(x) \leq C'' V_{\hat{\theta}(s)}(x)$ for any $(s, x) \in \mathcal{K} \times \mathbb{R}^N$, we get the result. □

LEMMA 6.5. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$ and $p \geq 1$ an integer. There exist $\epsilon > 0$ and $C > 0$ such that for any sequence $\epsilon = (\epsilon_k)_{k\geq 0}$ such that $\epsilon_k \leq \epsilon$ for k large enough, any sequence $\Delta = (\Delta_k)_{k\geq 0}$ and any $x \in \mathbb{R}^N$,*

$$\sup_{s\in\mathcal{K}} \sup_{k\geq 0} \mathbb{E}^{\Delta}_{x,\hat{\theta}(s)}[V^p(X_k) \mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}] \leq C V^p(x)\,.$$

Proof. Let K be a compact subset of Θ such that $\hat{\theta}(\mathcal{K}) \subset K$. We note in the sequel, $\theta_k = \hat{\theta}(s_k)$. We have for $k \geq 2$, using the Markov property and Lemma 6.2 and 6.4,

$$\begin{aligned}
&\mathbb{E}^{\Delta}_{x,\theta}[V^p_{\theta_{k-1}}(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}] \leq \mathbb{E}^{\Delta}_{x,\theta}[\Pi_{\theta_{k-1}} V^p_{\theta_{k-1}}(X_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}] \\
&\quad\leq \rho\left(\mathbb{E}^{\Delta}_{x,\theta}[V^p_{\theta_{k-2}}(X_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}] + \mathbb{E}^{\Delta}_{x,\theta}[(V^p_{\theta_{k-1}}(X_{k-1}) - V^p_{\theta_{k-2}}(X_{k-1}))\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}]\right) + C \\
&\qquad\leq \rho\left(\mathbb{E}^{\Delta}_{x,\theta}[V^p_{\theta_{k-2}}(X_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k-1}] + C'\epsilon_{k-1}\mathbb{E}^{\Delta}_{x,\theta}[V^p_{\theta_{k-2}}(X_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k-1}]\right) + C
\end{aligned}$$

so that by induction, we have

$$\mathbb{E}^{\Delta}_{x,\theta}[V^p_{\theta_{k-1}}(X_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\epsilon)\geq k}] \leq \prod_{l=1}^{k-1}(\rho(1+C'\epsilon_l))V^p_\theta(x) + \frac{C}{(1-\rho(1+C'\epsilon))}\,.$$

Choosing ϵ such that $\rho(1 + C'\epsilon) < 1$ and introducing again $0 \leq c_1 \leq c_2$ such that $c_1 V(x) \leq V_\theta(x) \leq c_2 V(x)$ for any $(x, \theta) \in \mathbb{R}^N \times K$, we get the result. □

This yields (**A3'(iii)**).

We now prove condition (**A3'(i)**).

Since $H_s(x) = S(x) - s$ with $S(x)$ at most quadratic in x, the choice of V directly ensures (5.2).

Considering (5.3): Since the Markov Chain satisfies the Drift condition (6.4), it is geometrically ergodic (see [11]), so there exist constants $0 < \gamma < 1$ and C such that

$$\|g_s\|_V = \|\sum_{k\geq 0}(\Pi^k_{\hat{\theta}(s)} H_s(x) - h(s))\|_V \leq \sum_{k\geq 0} C\gamma^k \|H_s\|_V < \infty\,.$$

Thus $\forall s \in \mathcal{S},\ g_s$ belongs to $\mathcal{L}_V$.

Thanks to (6.5), it is immediate that $\Pi_{\hat{\theta}(s)} g_s$ belongs to $\mathcal{L}_V$ too. This ends the proof of (**A3'(i)**).

We now move to the Hölder condition (**A3'(ii)**). We will use the following lemma:

LEMMA 6.6. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. For all $p \geq 1$ and any function $f \in \mathcal{L}_{V^p}$, $\forall (s,s') \in \mathcal{K}^2$ we have for $\theta = \hat\theta(s)$ and $\theta' = \hat\theta(s')$:*

$$\|\Pi_\theta f - \Pi_{\theta'} f\|_{V^{p+1/2}} \leq C_\mathcal{K} \|f\|_{V^{p+1/2}} \ |s - s'| \ .$$

Proof. For any $1 \leq j \leq N$ and $f \in \mathcal{L}_{V^p}$, we have

$$\Pi_{\theta,j} f(x) = (1 - r_{\theta,j}(x)) f(x) + \int_\mathbb{R} f(x_{j,b}) r_{\theta,j}(x,b) q(b|x_{-j},\theta) db \tag{6.13}$$

where $r_{\theta,j}(x) = \int_\mathbb{R} r_{\theta,j}(x,b) q(b|x_{-j},\theta) db$ is the average acceptance rate.

Let s and s' be two points in $\mathcal{K}$ and $s(\epsilon) = (1-\epsilon)s + \epsilon s'$ for $\epsilon \in [0,1]$ be a linear interpolation between s and s' (since $\mathcal{S}$ is convex, we can assume that $\mathcal{K}$ is a convex set so that $s(\epsilon) \in \mathcal{K}$ for any $\epsilon \in [0,1]$). We denote also $\theta(\epsilon) \triangleq \hat\theta(s(\epsilon))$ the associated path in Θ which is a C^1 function. To study the difference $|(\Pi_{\theta(1),j} - \Pi_{\theta(0),j}) f(x)|$, introduce $\Pi^1_{\theta,j} f(x) \triangleq (1 - r_{\theta,j}(x)) f(x)$ and $\Pi^2_{\theta,j} f(x) \triangleq \int_\mathbb{R} f(x_{j,b}) r_{\theta,j}(x,b) q(b|x_{-j},\theta) db$. We start with the difference $|(\Pi^2_{\theta(1),j} - \Pi^2_{\theta(0),j}) f(x)|$, and first note that under the conditional law $q(b|x_{-j},\theta)$, $b \sim \mathcal{N}(b_{\theta,j}(x), 1/|e_j|^2_\theta)$ where

$$b_{\theta,j}(x) \triangleq e_j^t p_{\theta,j}(x) = e_j^t x - \langle x, e_j \rangle_\theta / |e_j|^2_\theta \tag{6.14}$$

is the j-th coordinate of $p_{\theta,j}(x)$. We have

$$\Pi^2_{\theta,j} f(x) = \int_\mathbb{R} f(x_{j,0} + b e_j) r_{\theta,j}(x,b) \exp\left(- \frac{(b - b_{\theta,j}(x))^2 |e_j|^2_\theta}{2} \right) \frac{|e_j|_\theta}{\sqrt{2\pi}} db \,.$$

Since $r_{\theta,j}(x,b) = \tilde r_{\theta,j}(x,b) \wedge 1$ where $\tilde r_{\theta,j} \triangleq \frac{q(y|x_{j,b},\theta)}{q(y|x,\theta)}$ is a smooth function in θ, we have

$$|(\Pi^2_{\theta(1),j} - \Pi^2_{\theta(0),j}) f(x)| \leq \int_0^1 \int_\mathbb{R} f(x_{j,0} + b e_j) \left| \frac{d}{d\epsilon} \left(r_{\theta,j}(x,b) \exp(- \frac{(b - b_{\theta,j}(x))^2 |e_j|^2_\theta}{2}) \frac{|e_j|_\theta}{\sqrt{2\pi}} \right) \right| db \,.$$

However, one easily checks that there exists a constant $C_\mathcal{K}$ such that for any $s, s' \in \mathcal{K}$, ϵ, j and x (with $\theta = \theta(\epsilon)$):

$$\begin{aligned} &\left| \frac{d}{d\epsilon} \exp\left(- \frac{(b - b_{\theta,j}(x))^2 |e_j|^2_\theta}{2} \right) \frac{|e_j|_\theta}{\sqrt{2\pi}} \right| \\ &\quad \leq C_\mathcal{K} (1 + |b - b_{\theta,j}(x)|)^2 \exp\left(- \frac{(b - b_{\theta,j}(x))^2 |e_j|^2_\theta}{2} \right) \frac{|e_j|_\theta}{\sqrt{2\pi}} \left(\left| \frac{d}{d\epsilon} b_{\theta,j}(x) \right| + \left| \frac{d}{d\epsilon} |e_j|_\theta \right| \right) \ . \end{aligned} \tag{6.15}$$

Since $\frac{d}{d\epsilon}|e_j|_\theta = \frac{1}{2|e_j|_\theta} e_j^t \frac{d}{d\epsilon} \Gamma_\theta^{-1} e_j$, $\frac{d}{d\epsilon} \Gamma_\theta^{-1} = -\Gamma_\theta^{-1} \frac{d}{d\epsilon} \Gamma_\theta \Gamma_\theta^{-1}$ and $\frac{d}{d\epsilon} \Gamma_\theta = \frac{s_3' - s_3}{n + a_g}$ (see (3.14)), we deduce that there exists $C'_\mathcal{K}$ such that

$$\left| \frac{d}{d\epsilon} |e_j|_\theta \right| \leq C'_\mathcal{K} |s' - s| \,. \tag{6.16}$$

Similarly, updating the constant $C'_\mathcal{K}$, we have

$$\left| \frac{d}{d\epsilon} b_{\theta,j}(x) \right| \leq C'_\mathcal{K} (1 + |x|) |s' - s| \,. \tag{6.17}$$

Now, concerning the derivative of $\tilde r_{\theta,j}(x,b)$, since

$$\log(\tilde r_{\theta,j}(x,b)) = \frac{1}{2} \sum_{i=1}^n \left(|y_i - K_p^{\tilde\beta_i} \alpha|^2 - |y_i - K_p^{\beta_i} \alpha|^2 \right)$$

with $\tilde\beta_i = (x_{j,b})_{2k_g(i-1)+1}^{2k_g i}$, only one term of the previous sum is non zero. We deduce from the fact that K_p is bounded and from (3.14) that $|\frac{d}{d\epsilon}\log(\tilde r_{\theta,j}(x,b))| \le C|\frac{d}{d\epsilon}\alpha| \le C''_{\mathcal K}|s-s'|$, so that using the fact that $\tilde r_{\theta,j}(x,b)$ is uniformly bounded for $\theta \in \hat\theta(\mathcal K)$, $x \in \mathbb R^N$ and $b \in \mathbb R$, there exists a new constant $C''_{\mathcal K}$ such that

$$|\frac{d}{d\epsilon}\tilde r_{\theta,j}(x,b))| \le C''_{\mathcal K}|s-s'| \, . \tag{6.18}$$

Thus, using (6.16), (6.17) and (6.18), we get (for a new constant $C_{\mathcal K}$) that

$$\begin{aligned}&|\frac{d}{d\epsilon}\exp\left(-\frac{(b-b_{\theta,j}(x))^2|e_j|_\theta^2}{2}\right)\frac{|e_j|_\theta}{\sqrt{2\pi}}|\\ &\qquad \le C_{\mathcal K}(1+|x|)|s'-s|(1+|b-b_{\theta,j}(x)|)^2\exp\left(-\frac{(b-b_{\theta,j}(x))^2|e_j|_\theta^2}{2}\right)\frac{|e_j|_\theta}{\sqrt{2\pi}}\, .\end{aligned} \tag{6.19}$$

Since $|f(x)| \le \|f\|_{V^p}V^p(x)$ and $V(a+b) = (1+|a+b|^2) \le 2(V(a)+V(b))$, we get $|f(x_{0,j}+be_j)| \le C\|f\|_{V^p}(V^p(x_{0,j})+V^p(be_j))$ with $C = 2^{2p-1}$. Hence, there exists $C_{\mathcal K}$ such that for any $s, s' \in \mathcal K$, any j, x and $\epsilon \in [0,1]$ we have:

$$\begin{aligned}&\int_{\mathbb R}|f(x_{j,0}+be_j)|\left|\frac{d}{d\epsilon}\left(r_{\theta,j}(x,b)\exp\left(-\frac{(b-b_{\theta,j}(x))^2|e_j|_\theta^2}{2}\right)\frac{|e_j|_\theta}{\sqrt{2\pi}}\right)\right|db\\ &\qquad \le C_{\mathcal K}\|f\|_{V^p}V^p(x_{j,0})(1+|x|)|s'-s| \le C_{\mathcal K}\|f\|_{V^p}V^p(x)(1+|x|)|s'-s|\end{aligned} \tag{6.20}$$

where we have used the fact that a Gaussian variable has finite moments of all order. Since $(1+|x|) \le (2V(x))^{1/2}$, we get (updating $C_{\mathcal K}$) that

$$|(\Pi^2_{\theta(1),j}-\Pi^2_{\theta(0),j})f(x)| \le C_{\mathcal K}\|f\|_{V^p}V^{p+1/2}(x)|s'-s| \, . \tag{6.21}$$

Now, looking at the first term in (6.13), we deduce easily from the previous study for $f \equiv f(x)$ that

$$|(\Pi^1_{\theta(1),j}-\Pi^1_{\theta(0),j})f(x)| \le C_{\mathcal K}V(x)^{1/2}|s'-s||f(x)| \le C_{\mathcal K}\|f\|_{V^p}V^{p+1/2}(x)|s'-s| \tag{6.22}$$

so that adding (6.21) and (6.22), we get (updating again $C_{\mathcal K}$) that

$$\|(\Pi_{\theta(1),j}-\Pi_{\theta(0),j})f\|_{V^{p+1/2}} \le C_{\mathcal K}\|f\|_{V^p}|s'-s| \, . \tag{6.23}$$

We end the proof, saying that $\Pi_{\theta(1)}-\Pi_{\theta(0)} = \sum_{j=1}^N \Pi_{\theta(1),j+1,N}\circ(\Pi_{\theta(1),j}-\Pi_{\theta(0),j})\circ\Pi_{\theta(0),1,j-1}$ where $\Pi_{\theta,q,r} = \Pi_{\theta,r}\circ\Pi_{\theta,r-1}\circ\cdots\circ\Pi_{\theta,q}$ for any integer $q \le r$ and any $\theta \in \Theta$ so that using (6.11) and (6.23), we get the result.
□

Lemma 6.7. *Let $\mathcal K$ be a compact subset of $\mathcal S$. There exists a constant $C_{\mathcal K}$ such that for all $p \ge 1$ and any function $f \in \mathcal L_{V^p}$, $\forall (s,s') \in \mathcal K^2$, $\forall k \ge 0$, we have for $\theta = \hat\theta(s)$ and $\theta' = \hat\theta(s')$ that:*

$$\|\Pi_\theta^k f - \Pi_{\theta'}^k f\|_{V^{p+1/2}} \le C_{\mathcal K}\|f\|_{V^{p+1/2}}|s-s'| \, .$$

Proof. We use the same decomposition of the difference as previously:

$$\Pi_\theta^k f - \Pi_{\theta'}^k f = \sum_{i=1}^{k-1}\Pi_\theta^i(\Pi_\theta-\Pi_{\theta'})(\Pi_{\theta'}^{k-i-1}f-\pi_{\theta'}(f)) \, .$$

Using Lemma 6.6, Lemma 6.3 and the fact that $\|\Pi_\theta^k(f-\pi_\theta(f))\|_{V^p} \le \gamma^k\|f\|_{V^p}$ with $\gamma < 1$ (geometric ergodicity) we get:

$$\begin{aligned}\|\Pi_\theta^k f - \Pi_{\theta'}^k f\|_{V^{p+1/2}} &\le C\sum_{i=1}^{k-1}\|(\Pi_\theta-\Pi_{\theta'})(\Pi_{\theta'}^{k-i-1}f-\pi_{\theta'}(f))\|_{V^{p+1/2}}\\ &\le C\|f\|_{V^{p+1/2}}|s-s'|\sum_{i=1}^{k-1}\gamma^{k-i+1}\end{aligned}$$

and the lemma is proved. □

We now prove that h is a Lipschitz function, adapting linearly Appendix B in [3].

Let $x \in \mathbb{R}^N$ and denote $\theta = \hat{\theta}(s)$, $\theta' = \hat{\theta}(s')$. Write $h(s) - h(s') = A(s,s') + B(s,s') + C(s,s')$ where

$$\begin{aligned} A(s,s') &= (h(s) - \Pi_\theta^k H_s(x)) + (\Pi_{\theta'}^k H_{s'}(x) - h(s')) \\ B(s,s') &= \Pi_\theta^k H_s(x) - \Pi_{\theta'}^k H_s(x) \\ C(s,s') &= \Pi_{\theta'}^k H_s(x) - \Pi_{\theta'}^k H_{s'}(x) \ . \end{aligned}$$

Using the geometric ergodicity, Lemma 6.3 and Lemma 6.6, we get that there exists $C > 0$, independent of k such that:

$$\begin{aligned} |A(s,s')| &\le C\gamma^k \sup_{\mathcal{S}\in\mathcal{K}} \|H_s\|_V V(x) \\ |B(s,s')| &\le C \sup_{\mathcal{S}\in\mathcal{K}} \|H_s\|_V |s-s'| V^{3/2}(x) \\ |C(s,s')| &\le C \sup_{\mathcal{S}\in\mathcal{K}} \|H_s\|_V |s-s'| V(x) \ . \end{aligned}$$

This yields

$$|h(s) - h(s')| \le C V^{3/2}(x)(\gamma^k + |s-s'|) \ .$$

Hence, setting $k = [\log|s-s'|/\log(\gamma)]$ if $|s-s'| < 1$ and 1 otherwise, we get the result.

We can now end the proof of (**A3'(ii)**): On one hand we have:

$$\begin{aligned} |(\Pi_\theta^k H_s(x) - h(s)) - (\Pi_{\theta'}^k H_{s'}(x) - h(s'))| &\le |\Pi_\theta^k H_s(x) - \Pi_\theta^k H_{s'}(x)| \\ &+ |\Pi_\theta^k H_{s'}(x) - \Pi_{\theta'}^k H_{s'}(x)| + |h(s) - h(s')| \le C|s-s'|V^{3/2}(\beta_0) \ . \end{aligned}$$

On the other hand, we have thanks to the geometric ergodicity,

$$|(\Pi_\theta^k H_s(x) - h(s)) - (\Pi_{\theta'}^k H_{s'}(x) - h(s'))| \le C\gamma^k V^{3/2}(x) \ .$$

Hence for any t and $T \ge t$, we have

$$\begin{aligned} |\Pi_\theta^t g_s(x) - \Pi_{\theta'}^t g_{s'}(x)| \le \sum_{k=t}^{\infty} |(\Pi_\theta^k H_s(x) - h(s)) - (\Pi_{\theta'}^k H_{s'}(x) - h(s'))| \le \\ C V^{3/2}(x) \left[T|s-s'| + \frac{\gamma^{T+t}}{1-\gamma} \right] \ . \end{aligned}$$

Setting $T = [\log|s-s'|/\log(\gamma)]$ for $|s-s'| \le \delta < 1$ and $T = t$ otherwise, using also the fact that for any $0 < a < 1$ we have $|s-s'|\log|s-s'| = o(|s-s'|^a)$, we get the result.

This proves condition (**A3'(ii)**) for any $a < 1$.

6.4. (A4). This condition is not restrictive at all as we can set the step-size sequences as we need.

This concludes the demonstration of Theorem 3.1.

7. Conclusion and discussion. We have proposed a stochastic algorithm for Bayesian non-rigid deformable models building in the context of [1] as well as a proof of convergence toward a critical point of the observed likelihood. To our best knowledge, this is the first theoretical result of convergence for a well defined statistical point of view in the framework of deformable template.

The algorithm is based on a stochastic approximation of the EM algorithm based on a MCMC approximation of the posterior. If our main contribution concerns here mostly the theoretical side, the preliminary experiments presented here on the US-postal database shows that the stochastic approach can be easily implemented and is robust to noisy situations, giving better result than the previous deterministic schemes.

Many interesting questions remain open. One of them is the extension of the stochastic scheme to mixture of deformable models (defined as the multicomponents model in [1]) where the parameters are the weights of the individual components and for each component, the associated template and deformation law. This is of particular importance on real data analysis where the restriction to a unique deformable model could be too drastic. The design of such mixture corresponds to some kind of deformation invariant clustering approach of the data which is a basic issue in any unsupervised data analysis scheme. This extension is however not as straightforward as it could appear at first glance: due the high dimensional hidden deformation variables, a naive extension of Markovian dynamic to the label variables coding for the component value will have extremely poor mixing properties leading to unpractical algorithm. A less straightforward extension involving multiple MCMC chains is under study.

An other interesting extension is to consider diffeomorphic mapping and not only displacement fields for the hidden deformation. This appears to be particularly interesting in the context of Computational Anatomy where a one to one correspondence between the template and the observation is usually needed and cannot be guaranteed with linear spline interpolation schemes. This extension could be done in principle using tangent models based on geodesic shooting in the spirit of [13] but many numerical as well as theoretical works are still to be done on this side.

REFERENCES

[1] S. Allassonnière, Y. Amit, and A. Trouvé. Toward a coherent statistical framework for dense deformable template estimation. *Journal of the Royal Statistical Society*, 69:3–29, 2007.

[2] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable template. *Journal of the American Statistical Association*, 86(414):376–387, 1991.

[3] C. Andrieu, É. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312 (electronic), 2005.

[4] C. Chef d'Hotel, G. Hermosillo, and O. Faugeras. Variational methods for multimodal image matching. *International Journal of Computer Vision*, 50(3):329–343, 2002.

[5] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128, 1999.

[6] L. Dryden, I and V. Mardia, K. *Statistical Shape Analysis.* John Wiley and Sons, 1998.

[7] C. A. Glasbey and K. V. Mardia. A penalised likelihood approach to image warping. *Journal of the Royal Statistical Society, Series B*, 63:465–492, 2001.

[8] U. Grenander and M. I. Miller. Computational anatomy: an emerging discipline. *Quarterly of Applied Mathematics*, LVI(4):617–694, 1998.

[9] E. Kuhn and M. Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.*, 8:115–131 (electronic), 2004.

[10] S. Marsland, C. J. Twining, and C. J. Taylor. A minimum description length objective function for groupewise non-rigid image registration. *Image and Vision Computing*, 2007.

[11] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Communications and Control Engineering Series. Springer-Verlag London Ltd., London, 1993.

[12] C. Robert. *Méthodes de Monte Carlo par chaînes de Markov.* Statistique Mathématique et Probabilité. [Mathematical Statistics and Probability]. Éditions Économica, Paris, 1996.

[13] M. Vaillant, I. Miller, M, A. Trouvé, and L. Younes. Statistics on diffeomorphisms via tangent space representations. *Neuroimage*, 23(S1):S161–S169, 2004.