Bernoulli **16**(3), 2010, 641–678
DOI: 10.3150/09-BEJ229

Construction of Bayesian deformable models via a stochastic approximation algorithm: A convergence study

STÉPHANIE ALLASSONNIÈRE[1], ESTELLE KUHN[2] and ALAIN TROUVÉ[3]

[1]*CMAP Ecole Polytechnique, Route de Saclay, F-91128 Palaiseau, France.*
E-mail: Stephanie.Allassonniere@polytechnique.edu
[2]*LAGA, Université Paris 13, 99, Av. Jean-Baptiste Clément, F-93430 Villetaneuse and INRA, MIA, Domaine de Vilvert, F-78352 Jouy-en-Josas, France. E-mail: estelle.kuhn@jouy.inra.fr*
[3]*CMLA, ENS Cachan, CNRS, PRES UniverSud, 61 Av. Président Wilson, F-94230 Cachan, France.*
E-mail: Alain.Trouve@cmla.ens-cachan.fr

The problem of the definition and estimation of generative models based on deformable templates from raw data is of particular importance for modeling non-aligned data affected by various types of geometric variability. This is especially true in shape modeling in the computer vision community or in probabilistic atlas building in computational anatomy. A first coherent statistical framework modeling geometric variability as hidden variables was described in Allassonnière, Amit and Trouvé [*J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** (2007) 3–29]. The present paper gives a theoretical proof of convergence of effective stochastic approximation expectation strategies to estimate such models and shows the robustness of this approach against noise through numerical experiments in the context of handwritten digit modeling.

Keywords: Bayesian modeling; MAP estimation; non-rigid deformable templates; shape statistics; stochastic approximation algorithms

1. Introduction

In the field of image analysis, the statistical analysis and modeling of variable objects from a limited set of examples is still a quite challenging and largely unsolved problem, depending strongly on the use of adequate representations of data. One such representation is the so-called dense deformable template (DDT) framework (Amit, Grenander and Piccioni (1991)). Observations are defined as deformations, taken from a family of deformations of moderate "dimensionality", of a given exemplar or template. Such a representation appears particularly well adapted to the emerging field of computational anatomy, where one aims to build statistical models of the anatomical variability within a given population (Grenander and Miller (1998)). However, research on DDT has been mainly focused on the variational point of view, in which DDT is used as an efficient vehicle for a wide range of registration algorithms (Chef d'Hotel, Hermosillo and Faugeras (2002)). The problem of template estimation, viewed as a statistical estimation problem of parameters of generative models of images of deformable objects, has received much less attention.

1350-7265

In this paper, we consider the hierarchical Bayesian framework for dense deformable templates developed in Allassonnière, Amit and Trouvé (2007). 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. Individual deformations in this framework are treated as *hidden variables* (or, equivalently, as random effects in the mixed effects setting), whereas the template and the law of the deformations are parameters (or, equivalently, fixed effects) of interest. Parameter estimation for this model could be performed by maximum a posteriori (MAP), for which existence and consistency (as the number of observed images tends to infinity) has been proven (see Allassonnière, Amit and Trouvé (2007)). This contrasts with earlier work in Glasbey and Mardia (2001) using a penalized likelihood (PL), or the more recent maximum description length approach in Marsland, Twining and Taylor (2007), for which consistency cannot be proven because the deformations are considered as *nuisance parameters* to be estimated.

Our contribution in this paper is in defining effective and theoretically proven convergent stochastic algorithms for computing (local) maxima of the posterior on the parameters for Bayesian deformable template models. First, we specify an adapted stochastic approximation expectation minimization algorithm (SAEM algorithm) in this highly demanding framework where the hidden variables are non-rigid deformation fields living in finite- but high-dimensional space (typically hundreds or more dimensions). In particular, special attention must be paid to the sampling of the posterior distribution on the deformations. Obviously, MCMC samplers are unavoidable, but non-adaptive proposal distributions yielding simple symmetric random steps are of limited practical interest. The present paper introduces a more sophisticated hybrid Gibbs sampling scheme allowing an acceptable rejection rate during the estimation step. The overall algorithm is cast in the larger class of SAEM-MCMC algorithms introduced in Kuhn and Lavielle (2004). Second, we extend the convergence theory of SAEM-MCMC algorithms developed in Kuhn and Lavielle (2004) to cover the case of *unbounded* random effects arising naturally for deformation fields. The core material for this extension is based on the general stability and convergence results for stochastic algorithms with truncation on random boundaries given in Andrieu, Moulines and Priouret (2005). The main technical point is that in the presence of unbounded random effects and sequential estimation of the covariance matrix of the random effects, the usual regularity conditions for the solutions of the Poisson equations for the Markovian dynamic as a function of the parameters cannot be verified and have to be relaxed. As a result, we provide a new general stochastic approximation convergence theorem with a weaker set of assumptions. Third, we prove that the conditions for stability and convergence are fulfilled for our general SAEM-MCMC estimation algorithm for Bayesian dense deformable templates. Indeed, a well-known weakness of general stochastic approximation algorithm convergence results is that they rarely provide proofs of convergence for the algorithms *used in practice* since, in these implementations, the assumptions are not satisfied or are hard to verify (see Andrieu, Moulines and Priouret (2005)). Since stochastic approximation algorithms have recently started to attract interest in the field of deformable model estimation (see Allassonière et al. (2006) and Richard, Samson and Cuénod (2009)), our results provide the missing theoretical foundations and guidelines for their effective use. As an illustration of the potential of such SAEM-MCMC approaches in the context of deformable templates, particularly in the presence of noisy data, we present a set of experiments with images of handwritten digits.

This article is organized as follows. Section 2 briefly reviews the hierarchical Bayesian deformable template model proposed by Allassonnière, Amit and Trouvé (2007). In Section 3, we develop the SAEM-MCMC strategy for the estimation of the parameters. In Section 4, we then state our general convergence result for truncated stochastic approximation algorithms, extending the convergence theorem in Andrieu, Moulines and Priouret (2005), and state that members of the designed family of SAEM-MCMC algorithms in the previous section satisfy the assumptions. The proof of this last statement is postponed to Section 6, after Section 5 concentrates on experiments. In the final section, we provide a short discussion and conclusion.

2. Observation model

Let us recall the model introduced in Allassonnière, Amit and Trouvé (2007). We are given gray level images $(y_i)_{1\leq i\leq n}$ observed on a grid of pixels $\{v_u \in D \subset \mathbb{R}^2, u \in \Lambda\}$ which is embedded in a continuous domain $D \subset \mathbb{R}^2$ (typically, $D = [-1, 1] \times [-1, 1]$). Although the images are observed only at the pixels $(v_u)_u$, 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). Each observation y is assumed to be the discretization on a fixed pixel grid of a deformation of the template plus independent noise. Specifically, for each observation, there exists an *unobserved* deformation field $z : \mathbb{R}^2 \to \mathbb{R}^2$ such that for $u \in \Lambda$,

$$y(u) = I_0\big(v_u - z(v_u)\big) + \epsilon(u),$$

where ϵ denotes an independent additive noise.

2.1. Models for template and deformation

Our model takes into account two complementary aspects: photometry, indexed by p, and geometry, indexed by g. Estimating the template and the distribution on deformations directly as a continuous function would be an infinite-dimensional problem. We reduce this problem to a finite-dimensional one by restricting the search to a parameterized space of functions. The template $I_0 : \mathbb{R}^2 \to \mathbb{R}$ and the deformation $z : \mathbb{R}^2 \to \mathbb{R}^2$ are assumed to belong to fixed 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 centered at some fixed control points in the domain D: $(v_{p,j})_{1\leq j\leq k_p}$ and $(v_{g,j})_{1\leq j\leq k_g}$, respectively. They are therefore parameterized by the coefficients $\alpha \in \mathbb{R}^{k_p}$ and $\beta \in \mathbb{R}^{k_g} \times \mathbb{R}^{k_g}$, as follows. For all v in D, let

$$I_\alpha(v) \triangleq (\mathbf{K_p}\alpha)(v) \triangleq \sum_{j=1}^{k_p} K_p(v, v_{p,j})\alpha^j$$

and

$$z_\beta(v) \triangleq (\mathbf{K_g}\beta)(v) \triangleq \sum_{j=1}^{k_g} K_g(v, v_{g,j})\beta^j.$$

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

2.2. Parametric model

For clarity, we denote by $\mathbf{y}^t = (y_1^t, \ldots, y_n^t)$ and $\boldsymbol{\beta}^t = (\beta_1^t, \ldots, \beta_n^t)$ the collection of data and their corresponding deformation coefficients, respectively. The statistical model of the observations we consider is a generative hierarchical one. We assume conditional normal distributions for $\mathbf{y}$ and $\boldsymbol{\beta}$:

$$\begin{cases} \boldsymbol{\beta} \sim \bigotimes_{i=1}^{n} \mathcal{N}_{2k_g}(0, \Gamma_g) \Big| \Gamma_g, \\ \mathbf{y} \sim \bigotimes_{i=1}^{n} \mathcal{N}_{|\Lambda|}(z_{\beta_i} I_\alpha, \sigma^2 \mathrm{Id}) \Big| \boldsymbol{\beta}, \alpha, \sigma^2, \end{cases} \tag{1}$$

where $\bigotimes$ denotes the product of distributions of independent variables and $z I_\alpha(u) = I_\alpha(v_u - z(v_u))$, for u in Λ, denotes the action of the deformation on the template image. The parameters of interest are α (which determines the template image), σ^2 (the variance of the additive noise) and Γ_g (the covariance matrix of the variables β). We assume that $\theta = (\alpha, \sigma^2, \Gamma_g)$ belongs to an open parameter space Θ:

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

where $\|\cdot\|$ is the Euclidean norm, $\mathrm{Sym}^+_{2k_g}$ is the cone of real $2k_g \times 2k_g$ positive definite symmetric matrices and R is an arbitrary positive constant.

The likelihood of the observed data q_{obs} can be written as an integral over the unobserved deformation variables. Let us denote by q_{c} the conditional likelihood of the observations, given the hidden variables, and by q_{m} the likelihood of these missing variables. Then

$$q_{\mathrm{obs}}(\mathbf{y}|\theta) = \int q_{\mathrm{c}}(\mathbf{y}|\boldsymbol{\beta}, \alpha, \sigma^2) q_{\mathrm{m}}(\boldsymbol{\beta}|\Gamma_g)\, \mathrm{d}\boldsymbol{\beta},$$

where all of the densities are determined by the model (1).

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. By introducing prior distributions on the parameters, estimation with small samples is still possible. The regularizing effect of such priors can be seen in the parameter update steps (cf. Allassonnière, Amit and Trouvé (2007)). We use a generative model based on standard conjugate prior distributions for parameters $\theta = (\alpha, \sigma^2, \Gamma_g)$ with fixed hyper-parameters. Specifically, we assume a normal prior for α, an inverse Wishart prior on σ^2 and an inverse Wishart prior on Γ_g. Furthermore, all priors are assumed to

be independent. This yields $\theta = (\alpha, \sigma^2, \Gamma_g) \sim q_{\text{para}} \triangleq \nu_p \otimes \nu_g$, where

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

For two matrices A and B, we define $\langle A, B\rangle_F \triangleq \operatorname{tr}(A^t B)$, the Frobenius dot product on the set of matrices, where tr denotes the trace of the matrix.

3. Parameter estimation based on stochastic approximation EM

In our Bayesian framework, we obtain from Allassonnière, Amit and Trouvé (2007) the existence of the MAP estimator

$$\tilde{\theta}_n = \arg\max_{\theta \in \Theta} q_B(\theta|\mathbf{y}),$$

where q_B denotes the posterior likelihood of the parameters given the observations. The dependence on n refers to the sample size.

We now turn to the maximization problem for the penalized posterior distribution $q_B(\theta|\mathbf{y})$, which has no closed form in our case. Indeed, the probability density function is known up to a renormalization constant. That prevents a direct computation of $\tilde{\theta}_n$.

In order to solve this problem, we apply an "EM-like" algorithm to approximate the MAP estimator $\tilde{\theta}_n$. The solution we propose is to base our algorithm on the use of the stochastic approximation EM (SAEM). First, we outline certain characteristics of our model, which highlight the reasons for the choice of the particular procedure and enable us to simplify its implementation.

3.1. Model characteristics

An important characteristic of our model is that it belongs to the curved exponential family. In other words, the complete likelihood q can be written as

$$q(\mathbf{y}, \boldsymbol{\beta}, \theta) = \exp[-\psi(\theta) + \langle S(\boldsymbol{\beta}), \phi(\theta)\rangle],$$

where the sufficient statistic S is a Borel function on $\mathbb{R}^N$, with $N \triangleq 2nk_g$, taking its values in an open subset $\mathcal{S}$ of $\mathbb{R}^m$, and ψ, ϕ are two Borel functions on Θ. (Note that S, ϕ and ψ may also depend on $\mathbf{y}$, but since $\mathbf{y}$ will stay fixed in what follows, we omit this dependence.)

In our setting, we obtain the following formula:

$$\log q(\mathbf{y}, \boldsymbol{\beta}, \theta) = \log q_{\text{c}}(\mathbf{y}|\boldsymbol{\beta}, \theta) + \log q_{\text{m}}(\boldsymbol{\beta}|\theta) + \log q_{\text{para}}(\theta),$$

where q_{para} denotes the prior density of the parameters defined in the previous paragraph.

For any $1 \leq j \leq k_p$ and any $u \in \Lambda$, we denote by

$$K_p^{\beta}(u, j) = K_p\big(v_u - z_\beta(v_u), v_{p,j}\big)$$

the matrix which corresponds to the deformation of the kernel K_p through z_β at pixel u and evaluated at pixel location v_u. Then, for some constant C independent of θ,

$$\begin{aligned}\log q(\mathbf{y}, \boldsymbol{\beta}, \theta) = &\sum_{i=1}^{n}\left\{-\frac{|\Lambda|}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|y_i - K_p^{\beta_i}\alpha\|^2\right\}\\ &+ \sum_{i=1}^{n}\left\{-\frac{1}{2}\log(|\Gamma_g|) - \frac{1}{2}\beta_i^t\Gamma_g^{-1}\beta_i\right\}\\ &+ a_g\left\{-\frac{1}{2}\log(|\Gamma_g|) - \frac{1}{2}\langle\Gamma_g^{-1}, \Sigma_g\rangle_F\right\} - \frac{1}{2}(\alpha - \mu_p)^t\Sigma_p^{-1}(\alpha - \mu_p)\\ &+ a_p\left\{-\frac{1}{2}\log(\sigma^2) - \frac{\sigma_0^2}{2\sigma^2}\right\} + C.\end{aligned}$$

Note that $\|y_i - K_p^{\beta_i}\alpha\|^2 = (y_i - K_p^{\beta_i}\alpha)^t(y_i - K_p^{\beta_i}\alpha)$, where $K_p^{\beta_i}\alpha$ is another way to write the action of the deformation z_{β_i} on the template I_α, denoted previously by $z_{\beta_i} I_\alpha$. This form emphasizes the dot product between the sufficient statistics and a function of the parameters. It can be easily verified that the following matrix-valued functions are the sufficient statistics (up to a multiplicative constant):

$$\begin{aligned}S_1(\boldsymbol{\beta}) &= \sum_{1\leq i\leq n}(K_p^{\beta_i})^t y_i,\\ S_2(\boldsymbol{\beta}) &= \sum_{1\leq i\leq n}(K_p^{\beta_i})^t(K_p^{\beta_i}),\\ S_3(\boldsymbol{\beta}) &= \sum_{1\leq i\leq n}\beta_i^t\beta_i.\end{aligned}$$

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

$$\mathcal{S} = \{(S_1, S_2, S_3) \mid S_1 \in \mathbb{R}^{k_p}, S_2 + \sigma_0^2\Sigma_p^{-1} \in \text{Sym}_{k_p}^+, S_3 + a_g\Sigma_g \in \text{Sym}_{2k_g}^+\}.$$

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 $n_s = k_p + \frac{k_p(k_p+1)}{2} + k_g(2k_g + 1)$.

In Allassonnière, Amit and Trouvé (2007), the existence of the parameter estimate $\hat{\theta}(S)$ that maximizes the complete log-likelihood has been proven. It can easily be shown that α, σ^2 and Γ_g

are explicitly expressed with the above sufficient statistics as follows:

$$\begin{cases} \Gamma_g(S) = \dfrac{1}{n + a_g}(S_3 + a_g \Sigma_g), \\ \alpha(S) = \big(S_2 + \sigma^2(S)(\Sigma_p)^{-1}\big)^{-1}\big(S_1 + \sigma^2(S)(\Sigma_p)^{-1}\mu_p\big), \\ \sigma^2(S) = \dfrac{1}{n|\Lambda| + a_p}\big(n\|\mathbf{y}\|^2 + \alpha(S)^t S_2 \alpha(S) - 2\alpha(S)^t S_1 + a_p \sigma_0^2\big). \end{cases} \tag{3}$$

These formulae also prove the smoothness of $\hat{\theta}$ on the subset $\mathcal{S}$.

3.2. SAEM-MCMC algorithm with truncation on random boundaries

In order to compute the MAP estimator for our Bayesian model, we use a variant of the EM (expectation-maximization) algorithm from Dempster, Laird and Rubin (1977). This algorithm is quite natural when we have to maximize a likelihood under a hierarchical model with missing variables. Unfortunately, direct computation is not tractable and we have to find a solution to overcome the problematic E step where we have to compute an expectation with respect to the posterior distribution on $\boldsymbol{\beta}$ given $\mathbf{y}$. A first attempt was proposed in Allassonnière, Amit and Trouvé (2007), where this conditional distribution is approximated by a Dirac distribution at its mode (fast approximation with mode, or FAM-EM). The results are very interesting, but the authors point out the lack of convergence of the FAM-EM algorithm on a database with low signal-to-noise ratio (SNR). This is the issue we consider here. We propose an algorithm that ensures the convergence of the resulting sequence of estimators toward the MAP, whatever the quality of the input.

This solution is a procedure combining the stochastic approximation EM (SAEM) with Markov chain Monte Carlo (MCMC) in a more general framework than that proposed by Kuhn and Lavielle (2004), which, in turn, generalized the algorithm introduced by Delyon, Lavielle and Moulines (1999). Indeed, the kth iteration of the SAEM-MCMC algorithm consists of the following three steps.

Step 1: *Simulation step*. The missing data, that is, the deformation parameters $\boldsymbol{\beta}$, are drawn using the transition probability of a convergent Markov chain $\boldsymbol{\Pi}_\theta$ having the posterior distribution $q_{\text{post}}(\cdot|\mathbf{y}, \theta)$ as its stationary distribution:

$$\boldsymbol{\beta}_k \sim \boldsymbol{\Pi}_{\theta_{k-1}}(\boldsymbol{\beta}_{k-1}, \cdot).$$

Step 2: *Stochastic approximation step*. A stochastic approximation is performed on the complete log-likelihood using the simulated value of the missing data:

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

where $\boldsymbol{\Delta} = (\Delta_k)_k$ is a decreasing sequence of positive step-sizes.

Step 3: *Maximization step*. The parameters are updated in the M-step:

$$\theta_k = \arg\max_{\theta \in \Theta} Q_k(\theta).$$

The initial values Q_0 and θ_0 are arbitrarily chosen.

Remark 1. We cannot use the direct SAEM algorithm. Indeed, this would require sampling the hidden variable from the posterior distribution which is known only up to a normalization constant. This sampling is not possible here due to the complexity of the posterior probability density function.

Since our model belongs to the curved exponential family, the stochastic approximation step can easily be performed on the sufficient statistics S instead of on the complete log-likelihood. The maximization step (Step 3) is then straightforward, replacing in (3) the sufficient statistics with their corresponding stochastic approximations.

The convergence of this algorithm has been proven in Kuhn and Lavielle (2004) in the particular case of missing variables belonging to a compact subset of $\mathbb{R}^N$. However, as we set a Gaussian prior on the missing variables $\boldsymbol{\beta}$, we cannot assume that their support is compact. In order to provide an algorithm whose convergence can be proven in the current framework, we have to use a more general setting, introduced in Andrieu, Moulines and Priouret (2005), which involves truncation on random boundaries. The proof is given in Section 4. This can be formalized as follows.

Let $(\mathcal{K}_q)_{q\geq 0}$ be a sequence of increasing compact subsets of $\mathcal{S}$, such as $\bigcup_{q\geq 0}\mathcal{K}_q = \mathcal{S}$ and $\mathcal{K}_q \subset \text{int}(\mathcal{K}_{q+1})$, for all $q \geq 0$. Let $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k\geq 0}$ be a monotone non-increasing sequence of positive numbers and K a compact subset of $\mathbb{R}^N$. We construct a sequence $((\boldsymbol{\beta}_k, s_k))_{k\geq 0}$, as described in Algorithm 1, as follows. 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 these conditions is not satisfied, we reinitialize the sequences of $\boldsymbol{\beta}$ and s using a projection (for more details, see Andrieu, Moulines and Priouret (2005)), increase the size of the compact set and continue the iterations until convergence. This is detailed in the following steps.

Algorithm 1 Stochastic approximation with truncation on random boundaries

Set $\boldsymbol{\beta}_0 \in \text{K}$, $s_0 \in \mathcal{K}_0$, $\kappa_0 = 0$, $\zeta_0 = 0$ and $\nu_0 = 0$.
for all $k \geq 1$ **do**
  compute $\bar{s} = s_{k-1} + \Delta_{\zeta_{k-1}}(S(\bar{\boldsymbol{\beta}}) - s_{k-1})$
  where $\bar{\boldsymbol{\beta}}$ is sampled from a transition kernel $\boldsymbol{\Pi}_{\theta_{k-1}}(\boldsymbol{\beta}_{k-1}, \cdot)$
  if $\bar{s} \in \mathcal{K}_{\kappa_{k-1}}$ and $\|\bar{s} - s_{k-1}\| \leq \varepsilon_{\zeta_{k-1}}$ **then**
    set $(\boldsymbol{\beta}_k, s_k) = (\bar{\boldsymbol{\beta}}, \bar{s})$ and $\kappa_k = \kappa_{k-1}$, $\nu_k = \nu_{k-1} + 1$, $\zeta_k = \zeta_{k-1} + 1$
  else
    set $(\boldsymbol{\beta}_k, s_k) = (\tilde{\boldsymbol{\beta}}, \tilde{s}) \in \text{K} \times \mathcal{K}_0$ and $\kappa_k = \kappa_{k-1} + 1$, $\nu_k = 0$, $\zeta_k = \zeta_{k-1} + \phi(\nu_{k-1})$
    where $\phi : \mathbb{N} \to \mathbb{Z}$ is a function such that $\phi(k) > -k$ for any k
    and $(\tilde{\boldsymbol{\beta}}, \tilde{s})$ can be chosen through different ways (cf. (Andrieu, Moulines and Priouret, 2005)).
  end if
  $\theta_k = \hat{\theta}(s_k)$
end for

Initialization step: Initialize $\boldsymbol{\beta}_0$ and s_0 in two fixed compact sets K and $\mathcal{K}_0$, respectively.

Then, for the kth iteration, repeat the following four steps.

Step 1: *MCMC simulation step*. Draw one new element $\bar{\boldsymbol{\beta}}$ of the non-homogeneous Markov chain with respect to the kernel with the current parameters $\boldsymbol{\Pi}_{\theta_{k-1}}$ and starting at $\boldsymbol{\beta}_{k-1}$,

$$\bar{\boldsymbol{\beta}} \sim \boldsymbol{\Pi}_{\theta_{k-1}}(\boldsymbol{\beta}_{k-1}, \cdot).$$

Step 2: *Stochastic approximation step*. Compute

$$\bar{s} = s_{k-1} + \Delta_{\zeta_{k-1}}\big(S(\bar{\boldsymbol{\beta}}) - s_{k-1}\big). \tag{4}$$

Step 3: *Truncation on random boundaries*. If $\bar{s}$ is outside the current compact set $\mathcal{K}_{\kappa_{k-1}}$ or too far from the previous value s_k, then restart the stochastic approximation in the initial compact set, extend the truncation boundary to $\mathcal{K}_{\kappa_k}$ and start again with a bounded value of the missing variable. Otherwise, set $(\boldsymbol{\beta}_k, s_k) = (\bar{\boldsymbol{\beta}}, \bar{s})$ and keep the truncation boundary to $\mathcal{K}_{\kappa_{k-1}}$.

Step 4: *Maximization step*. Update the parameters using (3).

In this algorithm, the MCMC simulation step has to be explained since it involves the choice of the transition kernel of the Markov chain. Usually, one uses a Metropolis–Hastings algorithm in which a candidate value is sampled from a proposal distribution followed by an accept–reject step. However, there are different possible proposal distributions. The only requirement is that all of these kernels lead to an ergodic Markov chain whose stationary distribution is our posterior distribution. The choice among these possibilities should be based on the specific framework we are working in.

While minimizing the Kullback–Leibler distance between the stationary distribution $\boldsymbol{\beta} \to \boldsymbol{\pi}_\theta(\boldsymbol{\beta})$ and a tensorial product $\boldsymbol{\beta} \to \bigotimes_{i=1}^n p(\beta_i)$ corresponding to independent identically distributed missing variables, we get that p is proportional to $\frac{1}{n}\sum_{i=1}^n q_{\text{post}}(\cdot|y_i, \theta)$. As n tends to ∞ and for a given θ, p converges a.s. toward the prior pdf on the missing variable $q_{\text{m}}(\cdot|\theta)$. This suggests using as proposal the prior distribution which involves the current parameters.

On the other hand, the setting we are considering in this paper deals with high-dimensional missing variables. This raises several issues. If we simulate candidates for the hidden variable as a complete vector, it appears that most of the candidates are rejected. This is a typical high-dimensional concentration phenomenon: locally around a current point, the proportion of the space occupied by acceptable moves becomes negligible when the space dimension grows. From a more practical point of view, even if the proposed candidate is drawn with respect to the current prior distribution, it creates a deformation that is very different from the current one and too large for the corresponding deformed template to fit the observations. This yields very few possible moves from the current missing variable value and the algorithm is stuck in a non-optimal location or converges very slowly.

One solution is to update the chain one coordinate at a time, conditionally on the others. This corresponds to a Gibbs sampler and leads to more relevant candidates which have a higher chance of being accepted (cf. Amit (1996)). From an image analysis point of view, this puts stronger conditions on the kinds of deformations which are produced when proposing a candidate for

each coordinate. Knowing the tendency of the movement given by the other coordinates, the candidate will either confirm it or not, depending on whether this is a suitable movement. It will thus be accepted with a corresponding probability. Even if some coordinates remain unchanged, some others are updated, which enables the algorithm to visit a larger part of the missing variable support.

Remark 2. The index κ denotes the current active truncation set, the index ζ is the current index in the sequences $\boldsymbol{\Delta}$, $\boldsymbol{\varepsilon}$ and the index ν denotes the number of iterations since the last projection.

3.3. Transition probability of the Markov chain

We now explain how to simulate the missing variables by means of a Markov chain Monte Carlo algorithm having the posterior distribution as its stationary distribution. Due to the inherent high dimensionality N of $\boldsymbol{\beta}$, we consider a Gibbs sampler to sequentially scan all coordinates $\boldsymbol{\beta}^j$ for $1 \leq j \leq N$.

We define $\boldsymbol{\beta}^{-j} = (\boldsymbol{\beta}^l)_{l \neq j}$. We consider here a hybrid Gibbs sampler, that is, each step of the Gibbs sampler includes a Metropolis–Hastings step. The proposal law is chosen as $q_j(\cdot|\boldsymbol{\beta}^{-j}, \theta)$, that is, the conditional law based on the current parameter value θ derived from the normal distribution q_{m}.

If b is a proposed value at coordinate j, then the acceptance rate of the Metropolis–Hastings algorithm is given by

$$r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) = \left[\frac{q_j(b|\boldsymbol{\beta}^{-j}, \mathbf{y}, \theta) q_j(\boldsymbol{\beta}^j|\boldsymbol{\beta}^{-j}, \theta)}{q_j(\boldsymbol{\beta}^j|\boldsymbol{\beta}^{-j}, \mathbf{y}, \theta) q_j(b|\boldsymbol{\beta}^{-j}, \theta)} \wedge 1 \right].$$

Since

$$q_j(\boldsymbol{\beta}^j|\boldsymbol{\beta}^{-j}, \mathbf{y}, \theta) \propto q_{\mathrm{obs}}(\mathbf{y}|\boldsymbol{\beta}, \theta) q_j(\boldsymbol{\beta}^j|\boldsymbol{\beta}^{-j}, \theta),$$

the acceptance rate can be simplified to

$$r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) = \left[\frac{q_{\mathrm{obs}}(\mathbf{y}|\boldsymbol{\beta}_{b \to j}, \theta)}{q_{\mathrm{obs}}(\mathbf{y}|\boldsymbol{\beta}, \theta)} \wedge 1 \right],$$

where, for any $b \in \mathbb{R}$ and $1 \leq j \leq N$, we denote by $\boldsymbol{\beta}_{b \to j}$ the unique vector which is equal to $\boldsymbol{\beta}$ everywhere except at coordinate j, where it equals b. An illustration of the hybrid Gibbs sampler can be found in Robert (1996). The following steps are performed for each coordinate j.

Step 1: *Proposition*. Sample b with respect to the density $q_j(\cdot|\boldsymbol{\beta}^{-j}, \theta)$.

Step 2: *Accept–reject*. Compute $r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)$ and, with probability $r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)$, update $\boldsymbol{\beta}^j$ to b.

In Algorithm 2, we summarize the transition step of the Markov chain.

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

Require: $\boldsymbol{\beta} = \boldsymbol{\beta}_k$; $\theta = \theta_k$
  `Gibbs sampler:`
  for all $j = 1 : N$ **do**
    `Metropolis-Hastings procedure:`
    $b \sim q_j(\cdot|\boldsymbol{\beta}^{-j}, \theta)$;
    compute $r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) = [\frac{q_{\text{obs}}(\mathbf{y}|\boldsymbol{\beta}_{b\to j},\theta)}{q_{\text{obs}}(\mathbf{y}|\boldsymbol{\beta},\theta)} \wedge 1]$
    with probability $r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)$, update $\boldsymbol{\beta}^j$: $\boldsymbol{\beta}^j \leftarrow b$
  end for

This yields the transition probability kernel of our Markov chain on $\boldsymbol{\beta}$: for coordinate j, the kernel is

$$\begin{aligned}\boldsymbol{\Pi}_{\theta,j}(\boldsymbol{\beta}, \mathrm{d}\mathbf{z}) = &\left(\bigotimes_{m\neq j} \delta_{\boldsymbol{\beta}^m}(\mathrm{d}\mathbf{z}^m)\right) \\ &\times \Big[q_j(\mathrm{d}\mathbf{z}^j|\boldsymbol{\beta}^{-j}, \theta) r_j(\boldsymbol{\beta}^j, \mathrm{d}\mathbf{z}^j; \boldsymbol{\beta}^{-j}, \theta) \\ &\quad + \delta_{\boldsymbol{\beta}^j}(\mathrm{d}\mathbf{z}^j) \int \big(1 - r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)\big) q_j(b|\boldsymbol{\beta}^{-j}, \theta)\, \mathrm{d}b\Big]\end{aligned} \tag{5}$$

and $\Pi_\theta = \Pi_{\theta,N} \circ \cdots \circ \Pi_{\theta,1}$ is therefore the kernel associated with a complete scan.

4. Convergence analysis

We prove a general theorem on the convergence of stochastic approximations for which our algorithm convergence is a special case.

The hybrid Gibbs sampler used to generate the ergodic Markov chain does not satisfy some of the assumptions of the convergence result presented in Andrieu, Moulines and Priouret (2005). We therefore weaken some of their conditions, introducing an absorbing set for the stochastic approximation and weakening their Hölder conditions on some functions of the Markov chain.

4.1. Stochastic approximation convergence theorem

Let $\mathcal{S}$ be a subset of $\mathbb{R}^{n_s}$ for some integer n_s. Let X be a measurable space. For all $s \in \mathcal{S}$, let $H_s : X \to \mathcal{S}$ be a measurable function. Let $\boldsymbol{\Delta} = (\Delta_k)_k$ be a sequence of positive step-sizes.

Define the stochastic approximation sequence $(s_k)_k$ as follows:

$$\begin{cases} s_k = s_{k-1} + \Delta_{k-1} H_{s_{k-1}}(\boldsymbol{\beta}_k) & \text{with } \boldsymbol{\beta}_k \sim \boldsymbol{\Pi}_{s_{k-1}}(\boldsymbol{\beta}_{k-1}, \cdot), \quad \text{if } s_{k-1} \in \mathcal{S}, \\ s_k = s_c & \text{with } \boldsymbol{\beta}_k = \boldsymbol{\beta}_c, \quad \text{if } s_{k-1} \notin \mathcal{S}, \end{cases} \tag{6}$$

where $s_c \notin \mathcal{S}$, $\boldsymbol{\beta}_c \notin X$ and $(\boldsymbol{\Pi}_s)_{s\in\mathcal{S}}$ is a family of Markov transition probabilities on X. Denote by Q_Δ the transition which generates $((\boldsymbol{\beta}_k, s_k))_k$. We consider the natural filtration of the *non-homogeneous* chain $((\boldsymbol{\beta}_k, s_k))_k$ and denote respectively by $\mathbb{P}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}$ and $\mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}$ the probability measure and the corresponding expectation generated by this Markov chain starting at $(\boldsymbol{\beta}, s)$ and using the sequence $\boldsymbol{\Delta}$.

If the transition kernel $\boldsymbol{\Pi}_s$ of the Markov chain admits a stationary distribution π_s and if, for any $s \in \mathcal{S}$, H_s is integrable with respect to π_s, then we denote by h the mean field associated with our stochastic approximation so that

$$h(s) = \int H_s(\boldsymbol{\beta})\pi_s(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}.$$

The algorithm defined in (6) is usually designed to solve the equation $h(s) = 0$, where h is called the *mean field function*.

Let $(\mathcal{K}_q)_{q\geq 0}$ be a sequence of increasing compact subsets of $\mathcal{S}$ such as $\bigcup_{q\geq 0}\mathcal{K}_q = \mathcal{S}$ and $\mathcal{K}_q \subset \mathrm{int}(\mathcal{K}_{q+1}), \forall q \geq 0$. Let $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k\geq 0}$ be a monotone non-increasing sequence of positive numbers and K a subset of X.

Let $\Phi : X \times \mathcal{S} \to \mathrm{K} \times \mathcal{K}_0$ be a measurable function and $\phi : \mathbb{N} \to \mathbb{Z}$ be a function such that $\phi(k) > -k$ for any k. Define the *homogeneous* Markov chain

$$\big(Z_k = (\boldsymbol{\beta}_k, s_k, \kappa_k, \zeta_k, \nu_k)\big)_k \tag{7}$$

on $\mathcal{Z} \triangleq X \times \mathcal{S} \times \mathbb{N}^3$ with the following transition at iteration k:

- if $\nu_{k-1} = 0$, then draw $(\boldsymbol{\beta}_k, s_k) \sim Q_{\Delta_{\zeta_{k-1}}}(\Phi(\boldsymbol{\beta}_{k-1}, s_{k-1}), \cdot)$, otherwise draw $(\boldsymbol{\beta}_k, s_k) \sim Q_{\Delta_{\zeta_{k-1}}}((\boldsymbol{\beta}_{k-1}, s_{k-1}), \cdot)$;
- if $\|s_k - s_{k-1}\| \leq \varepsilon_{\zeta_{k-1}}$ and $s_k \in \mathcal{K}_{\kappa_{k-1}}$, then set $\kappa_k = \kappa_{k-1}$, $\zeta_k = \zeta_{k-1} + 1$ and $\nu_k = \nu_{k-1} + 1$, otherwise set $\kappa_k = \kappa_{k-1} + 1$, $\zeta_k = \zeta_{k-1} + \phi(\nu_{k-1})$ and $\nu_k = 0$.

Consider the following assumptions, generalized from Andrieu, Moulines and Priouret (2005). Define, for any $V : X \to [1, \infty]$ and any $g : X \to \mathbb{R}^{n_s}$, the norm

$$\|g\|_V = \sup_{\boldsymbol{\beta}\in X} \frac{\|g(\boldsymbol{\beta})\|}{V(\boldsymbol{\beta})}.$$

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 $w : \mathcal{S} \to [0, \infty[$ with the following properties:

(i) there exists an $M_0 > 0$ 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 exists a closed convex set $\mathcal{S}_a \subset \mathcal{S}$ for which $s \to s + \rho H_s(\boldsymbol{\beta}) \in \mathcal{S}_a$ for any $\rho \in [0, 1]$ and $(\boldsymbol{\beta}, s) \in X \times \mathcal{S}_a$ ($\mathcal{S}_a$ is absorbing), and such that for any $M_1 \in]M_0, \infty]$, the set $\mathcal{W}_{M_1} \cap \mathcal{S}_a$ 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 $s \in \mathcal{S} \setminus \mathcal{L}$ $\langle \nabla w(s), h(s)\rangle < 0$;

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

A2. For any $s \in \mathcal{S}$, the Markov kernel $\mathbf{\Pi}_s$ has a single stationary distribution π_s, $\pi_s \mathbf{\Pi}_s = \pi_s$. In addition, for all $s \in \mathcal{S}$, $H_s : X \to \mathcal{S}$ is measurable and $\int_X \|H_s(\boldsymbol{\beta})\| \pi_s(\mathrm{d}\boldsymbol{\beta}) < \infty$.

A3′. For any $s \in \mathcal{S}$, the Poisson equation $g - \mathbf{\Pi}_s g = H_s - \pi_s(H_s)$ has a solution g_s. There exist a function $V : X \to [1, \infty]$ such that $\{\boldsymbol{\beta} \in X, V(\boldsymbol{\beta}) < \infty\} \neq \varnothing$ and constants $a \in]0, 1]$, $q \geq 1$ and $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \mathcal{S}$:

(i)

$$\sup_{s \in \mathcal{K}} \|H_s\|_V < \infty, \tag{8}$$

$$\sup_{s \in \mathcal{K}} (\|g_s\|_V + \|\mathbf{\Pi}_s g_s\|_V) < \infty; \tag{9}$$

(ii)

$$\sup_{s, s' \in \mathcal{K}} \|s - s'\|^{-a} \{\|g_s - g_{s'}\|_{V^q} + \|\mathbf{\Pi}_s g_s - \mathbf{\Pi}_{s'} g_{s'}\|_{V^q}\} < \infty; \tag{10}$$

(iii) if k_0 is an integer, then there exist an $\bar{\varepsilon} > 0$ and a constant C such that for any sequence $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k \geq 0}$ satisfying $0 < \varepsilon_k \leq \bar{\varepsilon}$ for all $k \geq k_0$, for any sequence $\boldsymbol{\Delta} = (\Delta_k)_{k \geq 0}$ and for any $\boldsymbol{\beta} \in X$,

$$\sup_{s \in \mathcal{K}} \sup_{k \geq 0} \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}, s} \left[V^{pq}(\boldsymbol{\beta}_k) \mathbb{1}_{\sigma(\mathcal{K}) \wedge \nu(\boldsymbol{\varepsilon}) \geq k} \right] \leq C V^{pq}(\boldsymbol{\beta}), \tag{11}$$

where $\nu(\boldsymbol{\varepsilon}) = \inf\{k \geq 1, \|s_k - s_{k-1}\| \geq \varepsilon_k\}$, $\sigma(\mathcal{K}) = \inf\{k \geq 1, s_k \notin \mathcal{K}\}$ and the expectation is related to the non-homogeneous Markov chain $((\boldsymbol{\beta}_k, s_k))_{k \geq 0}$ using the step-size sequence $(\Delta_k)_{k \geq 0}$.

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

Theorem 1 (General convergence result for truncated stochastic approximation). *Assume* (A1′), (A2), (A3′) *and* (A4). *Let* $\mathrm{K} \subset X$ *be such that* $\sup_{\boldsymbol{\beta} \in \mathrm{K}} V(\boldsymbol{\beta}) < \infty$ *and* $\mathcal{K}_0 \subset \mathcal{W}_{M_0} \cap \mathcal{S}_a$ (*where* M_0 *is defined in* (A1′)) *and let* $(Z_k)_{k \geq 0}$ *be the sequence defined in equation* (7). *Then, for all* $\boldsymbol{\beta}_0 \in \mathrm{K}$ *and* $s_0 \in \mathcal{K}_0$, *we have* $\lim_{k \to \infty} d(s_k, \mathcal{L}) = 0$ $\bar{\mathbb{P}}_{\boldsymbol{\beta}_0, s_0, 0, 0, 0}$-*a.s*, *where* $\bar{\mathbb{P}}_{\boldsymbol{\beta}_0, s_0, 0, 0, 0}$ *is the probability measure associated with the chain* $(Z_k = (\boldsymbol{\beta}_k, s_k, \kappa_k, \zeta_k, \nu_k))_{k \geq 0}$ *starting at* $(\boldsymbol{\beta}_0, s_0, 0, 0, 0)$.

Proof. • The deterministic results obtained by Andrieu, Moulines and Priouret (2005) under their assumption (A1) remain true if we suppose the existence of an absorbing set, as defined in assumption (A1′). Indeed, the proofs in Andrieu, Moulines and Priouret (2005) can be carried through in the same way, restricting the sequences to the absorbing set. Therefore, we obtain the same properties. The first one (stated in Lemma 2.1 of Andrieu, Moulines and Priouret (2005)) gives the contraction property of the Lyapunov function w. We then have (as in Theorem 2.2 of Andrieu, Moulines and Priouret (2005)) the fact that a sequence of stochastic approximations stays almost surely in a compact set under some conditions on the perturbation. Finally, we establish the convergence of such a stochastic approximation.

• We then state a relation between the homogeneous and non-homogeneous chains, as done in Lemma 4.1 of Andrieu, Moulines and Priouret (2005).

• We now prove an equivalent version of Proposition 5.2 of Andrieu, Moulines and Priouret (2005), under our conditions. Indeed, the upper bound on the fluctuations of the noise sequence stated in this proposition is relaxed in our case, involving a different power on the function V.

Proposition 1. *Assume* (A3′). *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$ and let $\boldsymbol{\Delta} = (\Delta_k)_k$ and $\boldsymbol{\varepsilon} = (\varepsilon_k)_k$ be two non-increasing sequences of positive numbers such that $\lim_{k\to\infty} \varepsilon_k = 0$. Then, for p defined in* (A3′):

1. *there exists a constant C such that, for any $(\boldsymbol{\beta}, s) \in X \times \mathcal{K}$, any integer l and any $\delta > 0$,*

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

   where $S_{l,n}(\boldsymbol{\varepsilon}, \boldsymbol{\Delta}, \mathcal{K}) \triangleq \mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq n} \sum_{k=l}^{n} \Delta_k (H_{s_{k-1}}(\boldsymbol{\beta}_k) - h(s_{k-1}))$ and $\mathbb{P}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}$ is the probability measure generated by the non-homogeneous Markov chain $((\boldsymbol{\beta}_k, s_k))_k$ started from the initial condition $(\boldsymbol{\beta}, s)$;
2. *there exists a constant C such that for any $(\boldsymbol{\beta}, s) \in X \times \mathcal{K}$,*

$$\mathbb{P}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\big(\nu(\boldsymbol{\varepsilon}) < \sigma(\mathcal{K})\big) \leq C\left\{\sum_{k=l}^{\infty} (\Delta_k \varepsilon_k^{-1})^p\right\} V^{pq}(\boldsymbol{\beta}).$$

Proof. The proof of this proposition can proceed as in Andrieu, Moulines and Priouret (2005), except for the upper bound on the term involving the Hölder property (the second term in what follows). Under A3′(ii), this upper bound brings into play an exponent pq on the function V.

Indeed, rewrite $S_{1,n}(\boldsymbol{\varepsilon}, \boldsymbol{\Delta}, \mathcal{K})$ using the Poisson equation and decompose it into a sum of the following five terms:

$$T_n^{(1)} = \sum_{k=1}^{n} \Delta_k \big(g_{s_{k-1}}(\boldsymbol{\beta}_k) - \boldsymbol{\Pi}_{s_{k-1}} g_{s_{k-1}}(\boldsymbol{\beta}_{k-1})\big) \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k\}}, \tag{12}$$

$$T_n^{(2)} = \sum_{k=1}^{n-1} \Delta_{k+1} \big(\boldsymbol{\Pi}_{s_k} g_{s_k}(\boldsymbol{\beta}_k) - \boldsymbol{\Pi}_{s_{k-1}} g_{s_{k-1}}(\boldsymbol{\beta}_k)\big) \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k+1\}}, \tag{13}$$

$$T_n^{(3)} = \sum_{k=1}^{n-1} (\Delta_{k+1} - \Delta_k) \boldsymbol{\Pi}_{s_{k-1}} g_{s_{k-1}}(\boldsymbol{\beta}_k) \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k+1\}}, \tag{14}$$

$$T_n^{(4)} = \Delta_1 \boldsymbol{\Pi}_{s_0} g_{s_0}(\boldsymbol{\beta}_0) \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq 1\}} - \Delta_n \boldsymbol{\Pi}_{s_{n-1}} g_{s_{n-1}}(\boldsymbol{\beta}_n) \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq n\}}, \tag{15}$$

$$T_n^{(5)} = -\sum_{k=1}^{n-1} \Delta_k \boldsymbol{\Pi}_{s_{k-1}} g_{s_{k-1}}(\boldsymbol{\beta}_k) \mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})=k\}}. \tag{16}$$

We evaluate bounds for the first four quantities. Using the Minkowski inequality for $p/2 \geq 1$ and the Burkholder inequality (for $T_n^{(1)}$), we have

$$\sup_{s\in\mathcal{S}} \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\Big[\sup_{n\geq 0} \big\| T_n^{(1)} \big\|^p\Big] \leq C\left(\sum_{k=1}^{\infty} \Delta_k^2\right)^{p/2} \sup_{s\in\mathcal{S}} \sum_k \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\big[V^p(\boldsymbol{\beta}_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k\}}\big], \tag{17}$$

$$\sup_{s\in\mathcal{S}} \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\Big[\sup_{n\geq 0} \big\| T_n^{(2)} \big\|^p\Big] \leq C\left(\sum_{k=1}^{\infty} \Delta_k \varepsilon_k^{\alpha}\right)^{p} \sup_{s\in\mathcal{S}} \sum_k \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\big[V^{pq}(\boldsymbol{\beta}_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k\}}\big], \tag{18}$$

$$\sup_{s\in\mathcal{S}} \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\Big[\sup_{n\geq 0} \big\| T_n^{(3)} \big\|^p\Big] \leq C\Delta_1^p \sup_{s\in\mathcal{S}} \sum_k \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\big[V^p(\boldsymbol{\beta}_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k\}}\big], \tag{19}$$

$$\sup_{s\in\mathcal{S}} \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\Big[\sup_{n\geq 0} \big\| T_n^{(4)} \big\|^p\Big] \leq C\left(\sum_{k=1}^{\infty} \Delta_k^2\right)^{p/2} \sup_{s\in\mathcal{S}} \sum_k \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta}_0,s}\big[V^p(\boldsymbol{\beta}_k)\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k\}}\big], \tag{20}$$

where C is a constant which depends only on the compact set $\mathcal{K}$. The higher power pq appears because of the Hölder condition we assume on the solution of the Poisson equation.

Since, now, $T_n^{(5)}\mathbb{1}_{\{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq n\}} = 0$ and noting that $V(\boldsymbol{\beta}) \geq 1, \forall \boldsymbol{\beta} \in X$, we have $V(\boldsymbol{\beta})^p \leq V^{pq}(\boldsymbol{\beta})$. Successively applying (as in Andrieu, Moulines and Priouret (2005)) the Markov inequality, condition (11) and the Markov property to these upper bounds completes the proof of the first part of Proposition 1.

Concerning the second part, it follows from the same trick as above for upper-bounding the expectation of V^p by V^{pq}. This completes the proof of the proposition. □

It is now straightforward to prove the following proposition, which corresponds to Proposition 5.3 in Andrieu, Moulines and Priouret (2005).

Proposition 2. *Assume* (A3′) *and* (A4). *Then, for any subset* $\mathrm{K} \subset X$ *such that* $\sup_{\boldsymbol{\beta}\in\mathrm{K}} V(\boldsymbol{\beta}) < \infty$, *any* $M \in (M_0, M_1]$ *and any* $\delta > 0$, *we have* $\lim_{k\to\infty} A(\delta, \boldsymbol{\varepsilon}^{\leftarrow k}, M, \boldsymbol{\Delta}^{\leftarrow k}) = 0$, *where* $\boldsymbol{\varepsilon}^{\leftarrow k}$ *stands for the sequence* $\boldsymbol{\varepsilon}$ *delayed by* k *switches* ($\boldsymbol{\varepsilon}_l^{\leftarrow k} = \boldsymbol{\varepsilon}_{k+l}$ *for all* $l \in \mathbb{N}$) *and*

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

The convergence of the sequence $(s_k)_k$ follows from the proof of Theorem 5.5 of Andrieu, Moulines and Priouret (2005), which states the almost sure convergence due to the previous propositions. □

Remark 3. We can weaken the condition on p given in (A3′). Indeed, we can assume that (A3′) holds for any $p > 0$ provided that at least condition (11) is true also for a power equaled to 2 on V. This is needed in the proof when giving an upper bound for all of the T_n's using Jensen's inequality instead of Minkowski's inequality, as in Andrieu, Moulines and Priouret (2005). In this case, assumption (A4) would have to be satisfied for a power $\max(p, 2)$ instead of power 2.

4.2. Convergence theorem for dense deformable template model

We now give the convergence result for our estimation process which is an application of the previous theorem. In this section, we assume that σ^2 is fixed, which reduces θ to (α, Γ_g). In fact, due to the implicit definition of $\hat{\theta}$ given in equation (3), we were not able to prove the smoothness of the inverse of the function $s \mapsto \hat{\theta}(s)$, which is straightforward for fixed σ^2.

We can easily exhibit some of the functions involved in our procedure. Comparing equation (4) to equation (6), we have

$$H_s(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) - s. \tag{21}$$

Equation (3) gives the existence of the function $s \to \hat{\theta}(s)$. We denote by l the observed log-likelihood, $l(\theta) \triangleq \log \int q(\mathbf{y}, \boldsymbol{\beta}, \theta)\, \mathrm{d}\boldsymbol{\beta}$, and let $w(s) \triangleq -l \circ \hat{\theta}(s)$ and $h(s) \triangleq \int H_s(\boldsymbol{\beta}) q_{\mathrm{post}}(\boldsymbol{\beta}|\mathbf{y}, \hat{\theta}(s))\, \mathrm{d}\boldsymbol{\beta}$ for $s \in \mathcal{S}$.

Theorem 2. *The sequence of stochastic approximations $(s_k)_k$ related to the model defined in Section 2 and generated by Algorithms 1 and 2 satisfies assumptions* (A1′)(ii), (iii), (iv), (A2) *and* (A3′).

Proof. The details of the proof are given in Section 6. □

Corollary 1 (Convergence of dense deformable template building via stochastic approximation).

Assume that:

1. *there exist $p \geq 1$ and $a \in]0, 1[$ such that the sequences $\boldsymbol{\Delta} = (\Delta_k)_{k\geq 0}$ and $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k\geq 0}$ are non-increasing, positive and satisfy*

$$\sum_{k=0}^{\infty} \Delta_k = \infty, \qquad \lim_{k\to\infty} \varepsilon_k = 0 \quad \textit{and} \quad \sum_{k=1}^{\infty} \{\Delta_k^2 + \Delta_k \varepsilon_k^a + (\Delta_k \varepsilon_k^{-1})^p\} < \infty;$$

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 K *be a compact subset of $\mathbb{R}^N$ and $\mathcal{K}_0$ a compact subset of $S(\mathbb{R}^N)$.*

Let $(s_k)_{k\geq 0}$ and $(\theta_k)_{k\geq 0}$ be the two sequences defined in Algorithms 1 and 2. If we define $\mathcal{L}' \triangleq \{\theta \in \hat{\theta}(\mathcal{S}), \frac{\partial l}{\partial \theta}(\theta) = 0\}$, then $\hat{\theta}(\mathcal{L}) = \mathcal{L}'$ and

$$\lim_{k\to\infty} d(\theta_k, \mathcal{L}') = 0 \qquad \bar{\mathbb{P}}_{\boldsymbol{\beta}_0, s_0, 0, 0, 0}\text{-a.s.}$$

for all $\boldsymbol{\beta}_0 \in$ K and $s_0 \in \mathcal{K}_0$, where $\bar{\mathbb{P}}_{\boldsymbol{\beta}_0, s_0, 0, 0, 0}$ is the probability measure associated with the chain $(Z_k = (\boldsymbol{\beta}_k, s_k, \kappa_k, \zeta_k, \nu_k))_{k\geq 0}$ starting at $(\boldsymbol{\beta}_0, s_0, 0, 0, 0)$.

Proof. We first note that, as mentioned in Delyon, Lavielle and Moulines (1999) (Lemma 2, equation (36)), since $\hat{\theta}$, ϕ and ψ are smooth functions, it is easy to relate the convergence of

the stochastic approximation sequence $(s_k)_k$ to the convergence of the estimated parameter sequence $(\theta_k)_k$.

The proof then follows from the general stability result, Theorem 1, stated in Section 4.1 and from the previous Theorem 2. □

Remark 4. Note that condition (1) is easily checked for $\Delta_k = k^{-c}$ and $\varepsilon_k = k^{-c'}$ with $1/2 < c' < c < 1$. However, condition (2) has not yet been successfully proven and should be relaxed in future work.

5. Experiments

To illustrate our stochastic algorithm for the deformable template models, we consider handwritten digit images. For each digit class, we find the template, the corresponding noise variance and the geometric covariance matrices. (Note that in this experiment, the noise variance is no longer fixed and is estimated as the other parameters.) We use the United States Postal Service database, which contains a training set of around 7000 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 centered Gaussian noise to each image.

To be able to compare the results with the previous deterministic algorithm proposed in Allassonnière, Amit and Trouvé (2007), we use the same samples. In Figure 1 below, we show some of the training images.

A natural choice for the hyper-parameters on α and Γ_g is $\mu_p = 0$ and we induce the two covariance matrices Σ_p and Σ_g by the metric of the Hilbert spaces V_p and V_g (defined in Section 2.1) involving the correlation between the landmarks determined by the kernel. If we define the square matrices

$$\begin{aligned} M_p(k,k') &= K_p(v_{p,j}, v_{p,j'}) \qquad \forall 1 \le k,k' \le k_p, \\ M_g(k,k') &= K_g(v_{g,j}, v_{g,j'}) \qquad \forall 1 \le k,k' \le k_g, \end{aligned} \tag{22}$$

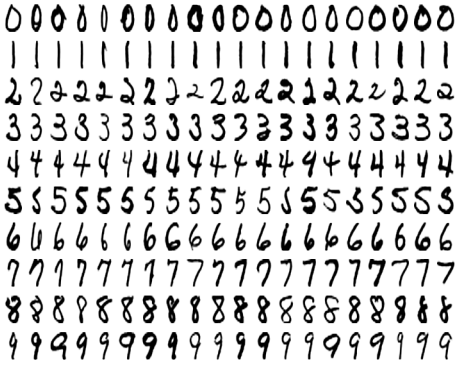

Figure 1. Some images from the training set used for the estimation of the model parameters (inverse video).

Figure 2. Estimated prototypes of digit 1 (20 images per class) for different hyper-parameters. Left: smoother geometry but larger photometric covariance in the spline kernel; right: more rigid geometry and smaller photometric covariance.

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 $\sigma_g = 0.3$. The deformation is computed in the $[-1, 1]^2$ square with $k_g = 6$ equidistributed landmarks on this domain. The template has been estimated with $k_p = 15$ equidistributed control points on $[-1.5, 1.5]^2$.

These two covariance matrices are important hyper-parameters; indeed, it has been shown in Allassonnière, Amit and Trouvé (2007) that changing the geometric covariance has an effect on the sharpness of the template images. As for the photometric hyper-parameter, it affects both the template and the geometry, in the sense that with a large variance, the kernel centered on one landmark spreads out to many of its neighbors. This leads to thicker shapes, as shown in the left panel of Figure 2. As a consequence, the template is biased: it is not "centered" in the sense that the mean of the deformations required to fit the data is not close to zero. For example, for digit "1", the main deformations should be 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 $z_{-\beta} = -z_\beta$, the deformations $(\mathrm{Id} + z_\beta)$ and $(\mathrm{Id} - z_\beta)$ have the same probability of being drawn under the estimated model. As shown on synthetic examples given in the left panel of Figure 3, there are many large dilated shapes. However, these examples were not in the training set and are not generated with other hyper-parameters (Figure 3, right panel). We have tried different relevant values and kept the best with regard to the visual results. In the following, we present 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 before exploring this particular region.

In the experiments presented here, the heating time lasts k_h (up to 150) iterations and the whole algorithm stops after, 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 have been reached. This yields

$$\Delta_k = \begin{cases} 1, & \forall 1 \le k \le k_h, \\ \dfrac{1}{(k - k_h)^d}, & \forall k > k_h \text{ for } d = 0.6 \text{ or } 1. \end{cases}$$

Figure 3. Synthetic examples corresponding to the two previous estimated templates of digit 1 (inverse video). Left: with a thicker shape; right: with a correct shape thickness.

To optimise the choice of the transition kernel $\mathbf{\Pi}_\theta$, we have run the algorithm with different kernels and compared the evolution of the simulated hidden variables, as well as the results on the estimated parameters. Some kernels, such as the ones mentioned above, do not yield good coverage of the infinite support of the unobserved variable. From this point of view, the hybrid Gibbs sampler we used has better properties and gives nice estimation results which are presented below.

5.1. Estimated template

We show here the results of the statistical learning algorithm for this model. Figure 4 shows two runs of the algorithm for a non-noisy database with 10 and 20 images per class. Ten images per class are enough to obtain satisfactory template images with high contrast.

Although it was proven in Allassonnière, Amit and Trouvé (2007) that the Kullback–Leibler divergence between $q(\cdot;\tilde{\theta})$ and the common density function for observations from a given class converges to its minimal value on the family $q(\cdot;\theta)$, we note that increasing the number of training images does not significantly improve the estimated photometric template. This apparently surprising fact can be explained as follows: since strong variations in appearance among the images may occur within a given class (consider, e.g., topological changes), the image distribution cannot be perfectly represented as a distribution around a single template. This distribution is better represented as being clustered around a major template and minor ones in a multimodal way. When the sample size is moderate, with a high probability, the sample basically contains images around the major mode and the parametric model fits these data quite accurately. When the sample size increases, the minor modes start to play a significant role as "outliers" with respect to the major mode in the data, resulting in a slightly more blurry template trying to accommodate the different modes. One way to overcome this fact is to use some clustering methods, as proposed in Allassonnière, Amit and Trouvé (2007). To visualize robustness with respect to the training set, we ran this algorithm with 20 images per class, randomly chosen from the whole database. The different runs are presented in Figure 5. The two left images show some templates which look like the ones obtained in the left panel of Figure 4 with the 20 first examples of the database. When outliers appear among the 20 randomly chosen training images, the template may become somewhat more blurry. This is observed for digits '2' and '4' (apparently the most variable digits) in the right panel of Figure 5. For digits where all of the images are close to each other (in term of deformation cost), the templates are stable.

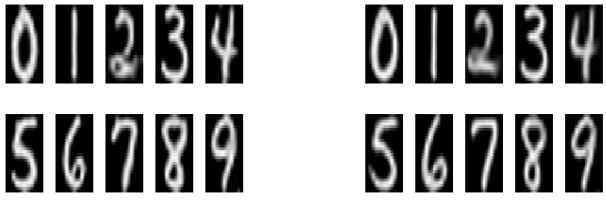

Figure 4. Estimated prototypes issued from left 10 images per class and right 20 images per class in the training set.

Figure 5. Templates estimated with randomly chosen samples from the whole United States Postal Service database. Each image is one run of the algorithm with the same initial conditions but different training sets of 20 images per digit each. The variability of the results is related to the huge variability within the USPS database.

The evolution of the template with the iterations can be viewed in Figure 6. The initialization of the template is the mean of the gray level images. As the iterations proceed, the templates become sharper. In particular, the estimated templates for digits with small geometric variability converge very fast. For digits like '2' or '4', where the geometric variability is higher, the convergence of the coupled parameters (photometry and geometry) is slowed down.

5.2. Photometric noise variance

The evolution of the noise variance along the SAEM-MCMC iterations is the same as the one observed with the "mode approximation EM" described in Allassonnière, Amit and Trouvé (2007). As shown in Figure 7, during the first iterations, the noise variance balances the inaccuracy of

Figure 6. Evolution of the templates with the algorithm iterations. Top line – left: mean gray level images of the 20 training samples; middle: template at the 50th iteration; right: template at the 100th iteration. Bottom line: template at the 150th iteration. The improvement is visible, very fast for some very simple shapes, such as the digit '1', and longer for very variable ones, such as the digit '2'. The higher geometric variability increases the fitting time of the algorithm.

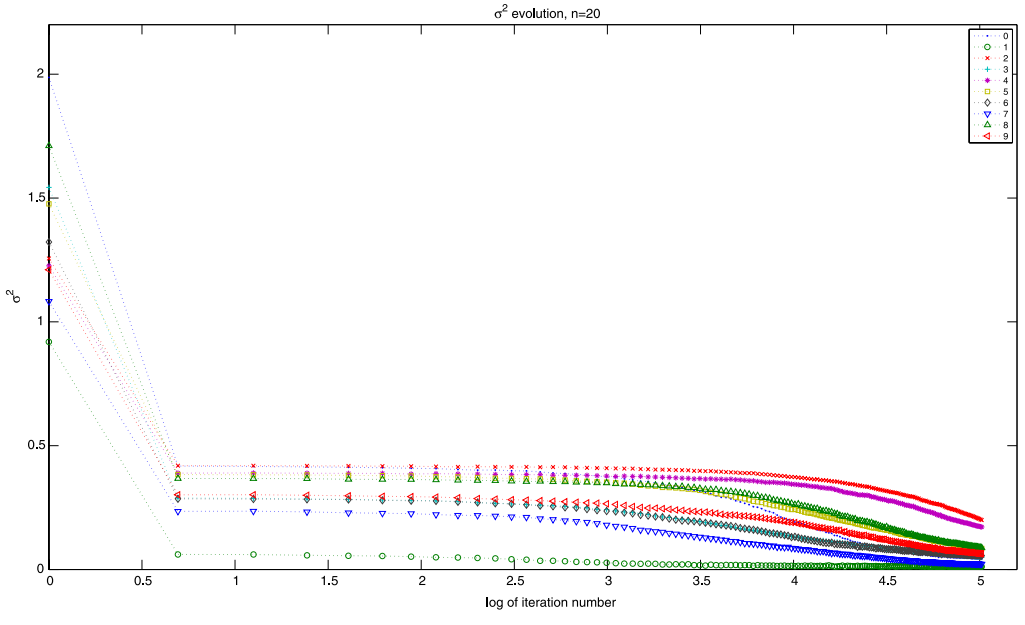

Figure 7. Evolution of the estimated noise variance using 20 images per class along the SAEM-MCMC algorithm. This confirms the visual effects seen on the templates: rapid convergence for some very constrained digits and slower convergence for the more variable ones.

the estimated template which is simply the gray level mean of the training set. As the iterations proceed, the template estimates become sharper, as does the estimate of the covariance matrix for the geometry. This yields very small residual noise. Note that, here, the final noise variance for the SAEM-MCMC algorithm, which is less than 0.1 for all digits, is less than the noise variance, which is between 0.2 and 0.3, for the mode approximation EM experimented in Allassonnière, Amit and Trouvé (2007) in the one component run. This can be explained by the stochastic nature of the algorithm, which enables it to escape from local minima provoking early terminations in the deterministic version.

5.3. Estimated geometric distribution

As mentioned previously, we have to fix the value of the hyper-parameter a_g of the prior on Γ_g. This quantity plays a significant role in the results. Indeed, to satisfy the theoretical conditions, we have to choose a_g larger than $4k_g + 1$, say $4 \times 36 + 1$, in our examples. From the geometry update equation, a barycenter between the 'sample' covariance and the prior, with the number n of images and a_g as coefficients, we find that the prior dominates when the training set is small. The covariance matrix stays close to the prior. Thus, we 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 learned geometry have to be satisfactory. This yields $a_g = 0.5$ or 0.1.

Figure 8. Effect of the estimated geometric distribution: 40 synthetic examples per class generated with the estimated parameters, 20 with the direct deformations and 20 with the symmetric deformations (inverse video).

As we have observed from Figure 8, parameter estimation is robust, regardless of whether the prior is degenerate or not. In addition, considering the update formulae, even if this law does not have a total weight equal to 1, it does not affect parameter estimation.

In Figure 8, we show a sample of some synthetic digits modeled by deformation templates drawn with the estimated parameters. Note that the resulting digits in Figure 8 look like some elements of the training set and seem to explain these data correctly, whereas the prior produces some non-relevant local deformations (cf. Figure 9). In particular, for some especially

Figure 9. Effect of the prior distribution on the deformation: 20 synthetic examples per class generated with the estimated template but the prior covariance matrix (inverse video).

geometrically constrained digits such as '0' or '1', the geometric variability reflects their constraints. For digits like the '2's, the training set is heterogeneous and shows a large geometric variability. When comparing to the deformations obtained by the mode approximation to EM in Allassonnière, Amit and Trouvé (2007), it seems that we here obtain a more variable 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 $\boldsymbol{\beta}$ 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). However, such a deformation can be proposed, leading to acceptance 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 stochastic EM algorithm.

5.4. Noise effect

As shown in Allassonnière, Amit and Trouvé (2007), in the presence of noise, the mode approximation algorithm does not converge toward the MAP estimator. In our setting, the consistency of the "SAEM-like" algorithm has been proven independently of the training set and thus noisy images can also be treated in the same way. These are the results we present here. Figure 10 shows two training examples per class for noise variance values $\sigma^2 = 1$ and $\sigma^2 = 2$. In Figures 11 and 12, 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. In contrast, the stochastic EM gives acceptable contrasted templates which look like those obtained in Figure 4. This becomes more significant as we increase the variance of the additive noise we introduce in the training set.

Concerning the choice of the hyper-parameters, it is not necessary to change all of them. For the photometric variance of the spline kernel, a small one could create some non-smooth templates and a large kernel would smooth the noise effect. However, we can keep the geometric

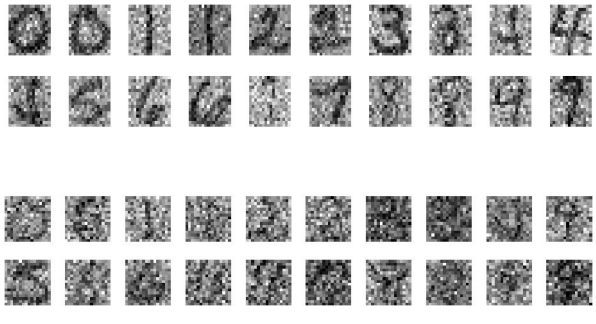

Figure 10. Two image examples per class of the noisy training set (variance – top: $\sigma^2 = 1$; bottom: $\sigma^2 = 2$).

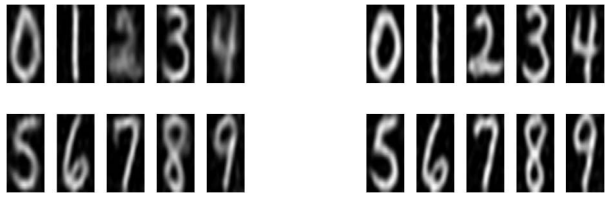

Figure 11. Estimated prototypes in a noisy setting $\sigma^2 = 1$. Left: with the mode approximation algorithm; right: with the SAEM-MCMC coupling procedure.

hyper-parameters unchanged. Here, we are presenting only experiments which seemed to provide a reasonable trade-off between these effects.

The geometry is also well estimated, despite the high level of noise in the training set. Figure 13 shows some synthetic examples, in which parameters are learned from the 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 learned on non-noisy images, even if some examples are not relevant. However, the global behavior has been learned.

The algorithm manages to catch the photometry (a contrasted and smoothed 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 template takes the longest time to converge and the estimate of σ^2 converges in a few iterations. In particular, the templates obtained in the left panel of Figure 4 with only 10 images per training digit set are obtained with a heating period of 25 iterations and 5 more steps with memory. The templates of the right panel of Figure 11 require 100 to 125 heating iterations in the 150 global 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 model.

6. Proof of Theorem 2

Here, we demonstrate Theorem 2, that is, that the stochastic approximation sequence satisfies assumptions (A1′)(ii), (iii), (iv), (A2) and (A3′).

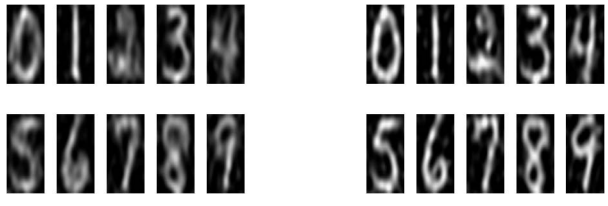

Figure 12. Estimated prototypes in a noisy setting $\sigma^2 = 2$. Left: with the mode approximation algorithm; right: with the SAEM-MCMC coupling procedure.

Figure 13. Effect of the noise on the geometric parameter estimation: 40 synthetic examples per class generated with the parameters estimated from the noisy training set (additive noise variance of 1, inverse video).

We recall that in this section, the parameter σ^2 is fixed so that $\theta = (\alpha, \Gamma)$. The sufficient statistic vector S, the set $\mathcal{S}$ and the explicit expression of $\hat{\theta}(s)$ have all been given in Section 4.2. As noted, $\hat{\theta}$ is a smooth function of $\mathcal{S}$.

We will prove that these conditions hold for any $p \geq 1$ and $a \in]0, 1[$.

6.1. Proof of assumption (A1′)

We recall the functions H, h and w, as in Delyon, Lavielle and Moulines (1999), are defined as follows:

$$H_s(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) - s,$$

$$h(s) = \int_{\mathbb{R}^N} H_s(\boldsymbol{\beta}) q_{\text{post}}(\boldsymbol{\beta}|\mathbf{y}, \hat{\theta}(s))\, \mathrm{d}\boldsymbol{\beta},$$

$$w(s) = -l(\hat{\theta}(s)).$$

As shown in Delyon, Lavielle and Moulines (1999), with these functions, we satisfy A1′(iii) and A1′(iv).

Moreover, since the interpolation kernel K_p is bounded, there exist $A > 0$ and $B \in \text{Sym}^+_{k_p}$ such that for any $\boldsymbol{\beta} \in \mathbb{R}^N$, we have

$$\|S_1(\boldsymbol{\beta})\| \leq A, \qquad 0 \leq S_2(\boldsymbol{\beta}) \leq B \quad \text{and} \quad 0 \leq S_3(\boldsymbol{\beta}),$$

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

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

$$\mathcal{S}_a \triangleq \{S \in \mathcal{S} \mid \|S_1\| \leq A, 0 \leq S_2 \leq B \text{ and } 0 \leq S_3\}.$$

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

$$s + \rho H_s(\boldsymbol{\beta}) \in \mathcal{S}_a \qquad \text{for any } \rho \in [0, 1], \text{ any } s \in \mathcal{S}_a \text{ and any } \boldsymbol{\beta} \in \mathbb{R}^N.$$

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

$$\mathcal{W}_M = \{s \in \mathcal{S}, w(s) \leq M\}.$$

On $\mathcal{S}_a$, s_1 and s_2 are bounded; writing $\hat{\theta}(s) = (\alpha(s), \Gamma(s))$, we deduce from (3) and from the boundedness of K_p that $\alpha(s)$ is bounded on $\mathcal{S}_a$ and $|y_i - K_p^{\beta_i}\alpha(s)|$ is uniformly bounded on $\beta_i \in \mathbb{R}^{2k_g}$ and $s \in \mathcal{S}_a$. Hence (recall that σ^2 is fixed here), there exists an $\eta > 0$ such that $q_{\mathrm{c}}(\mathbf{y}|\boldsymbol{\beta}, \hat{\theta}(s)) \geq \eta$ for any $s \in \mathcal{S}_a$ and $\boldsymbol{\beta} \in \mathbb{R}^N$. Thus,

$$w(s) \geq -\log\left(\int q_{\mathrm{m}}(\boldsymbol{\beta}, \hat{\theta}(s))\,\mathrm{d}\boldsymbol{\beta}\right) + \mathrm{C} \geq -\log(q_{\mathrm{para}}(\hat{\theta}(s))) + \mathrm{C} \geq -\log(q_{\mathrm{para}_{|\Gamma}}(\Gamma(s))) + \mathrm{C},$$

where C is a constant independent of $s \in S_a$. Since

$$-\log(q_{\mathrm{para}_{|\Gamma}}(\Gamma_g)) = \frac{a_g}{2}(\langle \Gamma_g^{-1}, \Sigma_g \rangle_F + \log|\Gamma_g|) \geq \frac{a_g}{2}\log|\Gamma_g|$$

and

$$\lim_{\|s\| \to +\infty, s \in \mathcal{S}_a} \log(|\Gamma_g(s)|) = \lim_{\|s\| \to +\infty, s \in \mathcal{S}_a} \log\big(|(s_3 + a_g\Sigma_g)/(n + a_g)|\big) = +\infty,$$

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).

6.2. Proof of assumption (A2)

We prove a classical sufficient condition (DRI1), used in Andrieu, Moulines and Priouret (2005), which will imply (A2) under the condition that H_s is dominated by V for any $s \in \mathcal{K}$.

(DRI1) For any $s \in \mathcal{S}$, $\mathbf{\Pi}_{\hat{\theta}(s)}$ is ϕ-irreducible and aperiodic. In addition, there exist a function $V : \mathbb{R}^N \to [1, \infty[$ and some $p \geq 2$ such that for any compact subset $\mathcal{K} \subset \mathcal{S}$, there exist an integer m and constants $0 < \lambda < 1$, $B > 0$, $\kappa > 0$, $\delta > 0$, a subset C of $\mathbb{R}^N$ and a probability measure ν such that

$$\sup_{s \in \mathcal{K}} \mathbf{\Pi}^m_{\hat{\theta}(s)} V^p(\boldsymbol{\beta}) \leq \lambda V^p(\boldsymbol{\beta}) + B \mathbb{1}_{\mathsf{C}}(\boldsymbol{\beta}), \tag{23}$$

$$\sup_{s \in \mathcal{K}} \mathbf{\Pi}_{\hat{\theta}(s)} V^p(\boldsymbol{\beta}) \leq \kappa V^p(\boldsymbol{\beta}) \qquad \forall \boldsymbol{\beta} \in \mathbb{R}^N, \tag{24}$$

$$\inf_{s \in \mathcal{K}} \mathbf{\Pi}^m_{\hat{\theta}(s)}(\boldsymbol{\beta}, A) \geq \delta \nu(A) \qquad \forall \boldsymbol{\beta} \in \mathsf{C}, \forall A \in \mathcal{B}(\mathbb{R}^N). \tag{25}$$

Remark 5. Note that condition (25) is equivalent to the existence of a small set C (defined below) which depends only on $\mathcal{K}$.

Notation 1. Let $(e_j)_{1 \leq j \leq N}$ be the canonical basis of $\mathbb{R}^N$. For any $1 \leq j \leq N$, let $E_{\theta,j} \triangleq \{\boldsymbol{\beta} \in \mathbb{R}^N \mid \langle \boldsymbol{\beta}, e_j \rangle_\theta = 0\}$ be the orthogonal space of $\mathrm{Span}\{e_j\}$ and $p_{\theta,j}$ be the orthogonal projection onto $E_{\theta,j}$, that is,

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

where $\langle \boldsymbol{\beta}, \boldsymbol{\beta}' \rangle_\theta = \sum_{i=1}^n \beta_i^t \Gamma_g^{-1} \beta_i'$ for $\theta = (\alpha, \Gamma_g)$ (i.e., the natural dot product associated with the covariance matrix Γ_g) and $\|\cdot\|_\theta$ is the corresponding norm.

For any $1 \leq j \leq N$ and $\theta \in \Theta$, we denote by $\mathbf{\Pi}_{\theta,j}$ the Markov kernel on $\mathbb{R}^N$ (5) associated with the Metropolis–Hastings step of the jth Gibbs sampler step on $\boldsymbol{\beta}$. We have $\mathbf{\Pi}_\theta = \mathbf{\Pi}_{\theta,N} \circ \cdots \circ \mathbf{\Pi}_{\theta,1}$.

We first recall the definition of a small set.

Definition 1 (cf. Meyn and Tweedie (1993)). *A set $\mathcal{E} \in \mathcal{B}(\mathcal{X})$ is called a* small set *for the kernel $\mathbf{\Pi}$ if there exist an $m > 0$ and a non-trivial measure ν_m on $\mathcal{B}(\mathcal{X})$ such that for all $\boldsymbol{\beta} \in \mathcal{E}$, $B \in \mathcal{B}(\mathcal{X})$,*

$$\mathbf{\Pi}^m(\boldsymbol{\beta}, B) \geq \nu_m(B). \tag{26}$$

*When (26) holds, we say that $\mathcal{E}$ is ν_m-*small.

We now prove the following lemma, which gives the existence of the small set C in (DRI1).

Lemma 1. *Let $\mathcal{E}$ be a compact subset of $\mathbb{R}^N$ and $\mathcal{K}$ a compact subset of $\mathcal{S}$. Then $\mathcal{E}$ is a small set of $\mathbb{R}^N$ for $\mathbf{\Pi}_{\hat{\theta}(s)}$, for any $s \in \mathcal{K}$.*

Proof. First, note that there exists an $a_{\mathrm{c}} > 0$ such that for any $\theta \in \Theta$, any $\boldsymbol{\beta} \in \mathbb{R}^N$ and any $b \in \mathbb{R}$, the acceptance rate $r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)$ is uniformly bounded below by a_{c} so that for any

$1 \leq j \leq N$ and any non-negative function f,

$$\begin{aligned}\mathbf{\Pi}_{\theta,j} f(\boldsymbol{\beta}) &\geq a_{\mathrm{c}} \int_{\mathbb{R}} f(\boldsymbol{\beta}^{-j} + b e_j) q_j(b|\boldsymbol{\beta}^{-j}, \theta)\,\mathrm{d}b \\ &= a_{\mathrm{c}} \int_{\mathbb{R}} f\big(p_{\theta,j}(\boldsymbol{\beta}) + z e_j/\|e_j\|_\theta\big) g_{0,1}(z)\,\mathrm{d}z,\end{aligned}$$

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

By induction, we have

$$\mathbf{\Pi}_\theta f(\boldsymbol{\beta}) \geq a_{\mathrm{c}}^N \int_{\mathbb{R}^N} f\left(p_{\theta,N,1}(\boldsymbol{\beta}) + \sum_{j=1}^N z_j\, p_{\theta,N,j+1}(e_j)/\|e_j\|_\theta\right) \prod_{j=1}^N g_{0,1}(z_j)\,\mathrm{d}z_j, \tag{27}$$

where $p_{\theta,q,r} = p_{\theta,r} \circ p_{\theta,r-1} \circ \cdots \circ p_{\theta,q}$ for any integers $q \leq r$ and $p_{\theta,N,N+1} = \mathrm{Id}$.

Let $A_\theta \in \mathcal{L}(\mathbb{R}^N)$ be the linear mapping on $\mathbb{R}^N$ defined by

$$A_\theta z = \sum_{j=1}^N z_j\, p_{\theta,N,j+1}(e_j)/\|e_j\|_\theta.$$

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

$$\int_{\mathbb{R}^N} f\big(p_{\theta,N,1}(\boldsymbol{\beta}) + A_\theta z_1^N\big) \prod_{j=1}^N g_{0,1}(z_j)\,\mathrm{d}z_j = \int_{\mathbb{R}^N} f(u) g_{p_{\theta,N,1}(\boldsymbol{\beta}), A_\theta A_\theta^t}(u)\,\mathrm{d}u,$$

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 exist two constants, $c_{\mathcal{K}} > 0$ and $C_{\mathcal{K}} > 0$, such that $c_{\mathcal{K}}\mathrm{Id} \leq A_\theta A_\theta^t \leq \mathrm{Id}/c_{\mathcal{K}}$ and $g_{p_{\theta,N,1}(\boldsymbol{\beta}), A_\theta A_\theta^t}(u) \geq C_{\mathcal{K}} g_{p_{\theta,N,1}(\boldsymbol{\beta}), \mathrm{Id}/c_{\mathcal{K}}}(u)$, uniformly for $\theta = \hat{\theta}(s)$ with $s \in \mathcal{K}$. Assuming that $\boldsymbol{\beta} \in \mathcal{E}$, since $\theta \to p_{\theta,N,1}$ is smooth and $\mathcal{E}$ is compact, we have $\sup_{\boldsymbol{\beta}\in\mathcal{E}, \theta=\hat{\theta}(s), s\in\mathcal{K}} \|p_{\theta,N,1}(\boldsymbol{\beta})\| < \infty$. Therefore, there exist $C'_{\mathcal{K}} > 0$ and $c'_{\mathcal{K}} > 0$ such that for any $(u, \boldsymbol{\beta}) \in \mathbb{R}^N \times \mathcal{E}$ and any $\theta = \hat{\theta}(s)$, $s \in \mathcal{K}$,

$$g_{p_{\theta,N,1}(\boldsymbol{\beta}), A_\theta A_\theta^t}(u) \geq C'_{\mathcal{K}} g_{0, \mathrm{Id}/c'_{\mathcal{K}}}(u). \tag{28}$$

Using (27) and (28), we deduce that for any A, for any $s \in \mathcal{K}$ and $\theta = \hat{\theta}(s)$,

$$\mathbf{\Pi}_\theta(\boldsymbol{\beta}, A) \geq C'_{\mathcal{K}} a_{\mathrm{c}}^N \nu_{\mathcal{K}}(A),$$

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

This yields the existence of the small set as well as equation (25). □

This property also implies the ϕ-irreducibility of the Markov chain $(\boldsymbol{\beta}_k)_k$ and its aperiodicity (cf. Meyn and Tweedie (1993), page 121).

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

$$V(\boldsymbol{\beta}) = 1 + \|\boldsymbol{\beta}\|^2. \tag{29}$$

In fact, we have the following property: $\exists C_{\mathcal{K}} > 0$ such that $\forall \boldsymbol{\beta} \in \mathbb{R}^N$,

$$\sup_{s\in\mathcal{K}} \|H_s(\boldsymbol{\beta})\| \leq C_{\mathcal{K}} V(\boldsymbol{\beta}).$$

This condition is required for the implication of A2 by DRI1.

We now prove condition (24).

Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$ and $p \geq 1$. For any $1 \leq j \leq N$, any $s \in \mathcal{K}$ and $\theta = \hat{\theta}(s)$, we have

$$\boldsymbol{\Pi}_{\theta,j} V^p(\boldsymbol{\beta}) \leq V^p(\boldsymbol{\beta}) + \int_{\mathbb{R}} V^p\big(p_{\theta,j}(\boldsymbol{\beta}) + z e_j/\|e_j\|_\theta\big) g_{0,1}(z)\,\mathrm{d}z.$$

Since $V(\boldsymbol{\beta}+h) \leq 2(V(\boldsymbol{\beta})+V(h))$ for any $\boldsymbol{\beta}, h \in \mathbb{R}^N$ and since there exist two constants, $c_{\mathcal{K}} > 0$ and $C_{\mathcal{K}} > 0$, such that for any $\boldsymbol{\beta} \in \mathbb{R}^N$, $\theta \in \hat{\theta}(\mathcal{K})$, $\|p_{\theta,j}(\boldsymbol{\beta})\| \leq C_{\mathcal{K}}\|\boldsymbol{\beta}\|$ and $\|e_j\|_\theta \geq 1/c_{\mathcal{K}}$, we have

$$\int_{\mathbb{R}} V^p\big(p_{\theta,j}(\boldsymbol{\beta}) + z e_j/\|e_j\|_\theta\big) g_{0,1}(z)\,\mathrm{d}z \leq 2^p C_{\mathcal{K}}^p V^p(\boldsymbol{\beta}) \int_{\mathbb{R}} \big(1 + V(c_{\mathcal{K}} z e_j)\big)^p g_{0,1}(z)\,\mathrm{d}z.$$

We deduce that there exists a $C'_{\mathcal{K}} > 0$ such that for any $\boldsymbol{\beta} \in \mathbb{R}^N$,

$$\sup_{\theta=\hat{\theta}(s), s\in\mathcal{K}} \boldsymbol{\Pi}_{\theta,j} V^p(\boldsymbol{\beta}) \leq C'_{\mathcal{K}} V^p(\boldsymbol{\beta}).$$

Then, by composition, $\boldsymbol{\Pi}_\theta V^p(\boldsymbol{\beta}) \leq C'^N_{\mathcal{K}} V^p(\boldsymbol{\beta})$ and (24) holds for any $p \geq 1$.

Now, consider the drift condition (23).

To prove this inequality, we prove the same inequality for a subsidiary function V_θ which depends on the parameters θ and then we deduce the result for V. So, let us define, for any $\theta = (\alpha, \Gamma_g)$, the function $V_\theta(\boldsymbol{\beta}) \triangleq 1 + \|\boldsymbol{\beta}\|^2_\theta$.

Lemma 2. *Let K be a compact subset of Θ. For any $p \geq 1$, there exist an $0 \leq \rho_K < 1$ and an $C_K > 0$ such that for any $\theta \in K$ and any $\boldsymbol{\beta} \in \mathbb{R}^N$, we have*

$$\boldsymbol{\Pi}_\theta V^p_\theta(\boldsymbol{\beta}) \leq \rho_K V^p_\theta(\boldsymbol{\beta}) + C_K.$$

Proof. The proposal distribution for $\boldsymbol{\Pi}_{\theta,j}$ is given by $q(\boldsymbol{\beta}|\boldsymbol{\beta}^{-j}, y, \theta) \overset{\text{law}}{=} p_{\theta,j}(\boldsymbol{\beta}) + z\frac{e_j}{\|e_j\|_\theta}$, where $z \sim \mathcal{N}(0,1)$. There then exists C_K such that for any $\boldsymbol{\beta} \in \mathbb{R}^N$ and any measurable set $A \in \mathcal{B}(\mathbb{R}^N)$,

$$\boldsymbol{\Pi}_{\theta,j}(\boldsymbol{\beta}, A) = (1 - a_{\theta,\boldsymbol{\beta}})\mathbb{1}_A(\boldsymbol{\beta}) + a_{\theta,\boldsymbol{\beta}} \int_{\mathbb{R}} \mathbb{1}_A\left(p_{\theta,j}(\boldsymbol{\beta}) + z\frac{e_j}{\|e_j\|_\theta}\right) g_{0,1}(z)\,\mathrm{d}z,$$

where $a_{\theta,\boldsymbol{\beta}} \geq a_{\mathrm{c}}$ (a_{c} is a lower bound for the acceptance rate).

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

$$\begin{aligned}\boldsymbol{\Pi}_{\theta,j} V_\theta^p(\boldsymbol{\beta}) &= (1 - a_{\theta,\boldsymbol{\beta}}) V_\theta^p(\boldsymbol{\beta}) + a_{\theta,\boldsymbol{\beta}} \int_{\mathbb{R}} \big(V_\theta(p_{\theta,j}(\boldsymbol{\beta})) + z^2\big)^p g_{0,1}(z)\,\mathrm{d}z \\ &\le (1 - a_{\theta,\boldsymbol{\beta}}) V_\theta^p(\boldsymbol{\beta}) \\ &\quad + a_{\theta,\boldsymbol{\beta}} \left(V_\theta^p(p_{\theta,j}(\boldsymbol{\beta})) + C_K V_\theta^{p-1}(p_{\theta,j}(\boldsymbol{\beta})) \int_{\mathbb{R}} (1 + z^2)^p g_{0,1}(z)\,\mathrm{d}z \right) \\ &\le (1 - a_{\theta,\boldsymbol{\beta}}) V_\theta^p(\boldsymbol{\beta}) + a_{\theta,\boldsymbol{\beta}} V_\theta^p(p_{\theta,j}(\boldsymbol{\beta})) + C'_K V_\theta^{p-1}(p_{\theta,j}(\boldsymbol{\beta})).\end{aligned}$$

In the last inequality, we have used the fact that a Gaussian variable has bounded moments of any order. Since $a_{\theta,\boldsymbol{\beta}} \ge a_{\mathrm{c}}$ and $\|p_{\theta,j}(\boldsymbol{\beta})\|_\theta \le \|\boldsymbol{\beta}\|_\theta$ ($p_{\theta,j}$ is an orthonormal projection for the dot product $\langle\cdot,\cdot\rangle_\theta$), we get that $\forall \eta > 0$, $\exists C_{K,\eta}$ such that $\forall \boldsymbol{\beta} \in \mathbb{R}^N$ and $\forall \theta \in K$,

$$\boldsymbol{\Pi}_{\theta,j} V_\theta^p(\boldsymbol{\beta}) \le (1 - a_{\mathrm{c}}) V_\theta^p(\boldsymbol{\beta}) + (a_{\mathrm{c}} + \eta) V_\theta^p(p_{\theta,j}(\boldsymbol{\beta})) + C_{K,\eta}.$$

By induction, we show that

$$\boldsymbol{\Pi}_\theta V_\theta^p(\boldsymbol{\beta}) \le \sum_{u \in \{0,1\}^N} \prod_{j=1}^N (1 - a_{\mathrm{c}})^{1-u_j} (a_{\mathrm{c}} + \eta)^{u_j} V_\theta^p(p_{\theta,u}(\boldsymbol{\beta})) + \frac{C_{K,\eta}}{\eta}\big((1 + \eta)^{N+1} - 1\big),$$

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 $p_\theta = p_{\theta,N} \circ \cdots \circ p_{\theta,1}$ and note that $p_{\theta,j}$ is contracting so that

$$\boldsymbol{\Pi}_\theta V_\theta^p(\boldsymbol{\beta}) \le b_{c,\eta} V_\theta^p(\boldsymbol{\beta}) + (a_{\mathrm{c}} + \eta)^N V_\theta^p(p_\theta(\boldsymbol{\beta})) + \frac{C_{K,\eta}}{\eta}\big((1 + \eta)^{N+1}\big),$$

for $b_{c,\eta} = (\sum_{u \in \{0,1\}^N, u \ne \mathbf{1}} \prod_{j=1}^N (1 - a_{\mathrm{c}})^{1-u_j} (a_{\mathrm{c}} + \eta)^{u_j})$.

To end the proof, we need to check that p_θ is strictly contracting uniformly on K. Indeed, $\|p_\theta(\boldsymbol{\beta})\|_\theta = \|\boldsymbol{\beta}\|_\theta$ implies that $p_{\theta,j}(\boldsymbol{\beta}) = \boldsymbol{\beta}$ for any $1 \le j \le N$. This yields $\langle \boldsymbol{\beta}, e_j\rangle_\theta = 0$ and thus $\boldsymbol{\beta} = 0$ since $(e_j)_{1 \le j \le N}$ 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 $\|p_\theta(\boldsymbol{\beta})\|_\theta \le \rho_K \|\boldsymbol{\beta}\|_\theta$ for any $\boldsymbol{\beta}$ and $\theta \in K$. Changing ρ_K for $1 > \rho'_K > \rho_K$, we get $(1 + \rho_K^2 \|\boldsymbol{\beta}\|_\theta^2)^p \le \rho'^{2p}_K (1 + \|\boldsymbol{\beta}\|_\theta^2)^p + C''_K$ for some uniform constant C''_K. Therefore,

$$\boldsymbol{\Pi}_\theta V_\theta^p(\boldsymbol{\beta}) \le b_{c,\eta} V_\theta^p(\boldsymbol{\beta}) + \rho'^{2p}_K (a_{\mathrm{c}} + \eta)^N V_\theta^p(\boldsymbol{\beta}) + C''_{K,\eta}.$$

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

Next, we prove the expected inequality for the function V.

Lemma 3. *For any compact set $K \subset \Theta$ and any $p \ge 1$, there exist $0 < \rho_K < 1$, $C_K > 0$ and m_0 such that $\forall m \ge m_0$, $\forall \theta \in K$, $\forall \boldsymbol{\beta} \in \mathbb{R}^N$,*

$$\boldsymbol{\Pi}_\theta^m V^p(\boldsymbol{\beta}) \le \rho_K V^p(\boldsymbol{\beta}) + C_K.$$

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

This completes the proof of (23) and, at the same time, of A2.

6.3. Proof of assumption A3′

The geometric ergodicity of the Markov chain, implied by the drift condition (23), ensures the existence of a solution of the Poisson equation (cf. Meyn and Tweedie (1993)):

$$g_{\hat\theta(s)}(\boldsymbol{\beta}) = \sum_{k\ge 0} \bigl(\boldsymbol{\Pi}_{\hat\theta(s)}^k H_s(\boldsymbol{\beta}) - h(s)\bigr).$$

We first prove condition A3′(i).

Since $H_s(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) - s$ with $S(\boldsymbol{\beta})$ at most quadratic in $\boldsymbol{\beta}$, the choice of V directly ensures (8).

Due to the result presented in Douc, Moulines and Rosenthal (2004), there exist upper bounds for the convergence rates and the constants involved in the quantification of the geometric ergodicity of all of the chains indexed by $s \in \mathcal{K}$ which only depend on m, λ, B, δ. Therefore, these constants only depend on the fixed compact set $\mathcal{K}$. This yields the uniform ergodicity of the family of Markov chains on $\mathcal{K}$. Therefore, there exist constants $0 < \gamma_{\mathcal{K}} < 1$ and $C_{\mathcal{K}} > 0$ such that

$$\|g_{\hat\theta(s)}\|_V = \left\| \sum_{k\ge 0} \bigl(\boldsymbol{\Pi}_{\hat\theta(s)}^k H_s(\boldsymbol{\beta}) - h(s)\bigr) \right\|_V \le \sum_{k\ge 0} C_{\mathcal{K}} \gamma_{\mathcal{K}}^k \|H_s\|_V < \infty.$$

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

Repeating the same calculation as above, it is immediate that $\boldsymbol{\Pi}_{\hat\theta(s)} g_{\hat\theta(s)}$ also belongs to $\mathcal{L}_V$. This completes the proof of A3′(i).

We now move to the Hölder condition A3′(ii). We will use the following lemmas which state Lipschitz conditions on the transition kernel and its iterates.

Lemma 4. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. There exists a constant $C_{\mathcal{K}}$ such that for any $p \ge 1$ and any function $f \in \mathcal{L}_{V^p}$, $\forall (s, s') \in \mathcal{K}^2$, we have*

$$\bigl\| \boldsymbol{\Pi}_{\hat\theta(s)} f - \boldsymbol{\Pi}_{\hat\theta(s')} f \bigr\|_{V^{p+1/2}} \le C_{\mathcal{K}} \|f\|_{V^p} \|s - s'\|.$$

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

$$\boldsymbol{\Pi}_{\theta,j} f(\boldsymbol{\beta}) = \bigl(1 - r_j(\boldsymbol{\beta}, \theta)\bigr) f(\boldsymbol{\beta}) + \int_{\mathbb{R}} f(\boldsymbol{\beta}_{b\to j}) r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) q_j(b|\boldsymbol{\beta}^{-j}, \theta)\, \mathrm{d}b,$$

where $r_j(\boldsymbol{\beta}, \theta) = \int_{\mathbb{R}} r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) q_j(b|\boldsymbol{\beta}^{-j}, \theta)\, \mathrm{d}b$ 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 also denote by $\theta(\epsilon) \triangleq \hat{\theta}(s(\epsilon))$ the associated path in Θ which is a continuously differentiable function. To study the difference $\|(\mathbf{\Pi}_{\theta(1),j} - \mathbf{\Pi}_{\theta(0),j}) f(\boldsymbol{\beta})\|$, introduce $\mathbf{\Pi}^1_{\theta,j} f(\boldsymbol{\beta}) \triangleq (1 - r_j(\boldsymbol{\beta},\theta)) f(\boldsymbol{\beta})$ and $\mathbf{\Pi}^2_{\theta,j} f(\boldsymbol{\beta}) \triangleq \int_{\mathbb{R}} f(\boldsymbol{\beta}_{b\to j}) \times r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) q_j(b|\boldsymbol{\beta}^{-j}, \theta)\,\mathrm{d}b$. We start with the difference $\|(\mathbf{\Pi}^2_{\theta(1),j} - \mathbf{\Pi}^2_{\theta(0),j}) f(\boldsymbol{\beta})\|$. First, note that under the conditional law $q_j(b|\boldsymbol{\beta}^{-j}, \theta)$, $b \sim \mathcal{N}(b_{\theta,j}(\boldsymbol{\beta}), 1/\|e_j\|^2_\theta)$, where

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

is the jth coordinate of $p_{\theta,j}(\boldsymbol{\beta})$. We have

$$\mathbf{\Pi}^2_{\theta,j} f(\boldsymbol{\beta}) = \int_{\mathbb{R}} f(\boldsymbol{\beta}_{0\to j} + b e_j) r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) \exp\left(-\frac{(b - b_{\theta,j}(\boldsymbol{\beta}))^2 \|e_j\|^2_\theta}{2}\right) \frac{\|e_j\|_\theta}{\sqrt{2\pi}}\,\mathrm{d}b.$$

Since $r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) = \tilde{r}_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) \wedge 1$, where $\tilde{r}_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) \triangleq \frac{q_{\mathrm{obs}}(\mathbf{y}|\boldsymbol{\beta}_{b\to j}, \theta)}{q_{\mathrm{obs}}(\mathbf{y}|\boldsymbol{\beta}, \theta)}$ is a smooth function in θ, we have

$$\begin{aligned}
&\left\|\left(\mathbf{\Pi}^2_{\theta(1),j} - \mathbf{\Pi}^2_{\theta(0),j}\right) f(\boldsymbol{\beta})\right\| \\
&\quad\le \int_0^1 \int_{\mathbb{R}} \|f(\boldsymbol{\beta}_{0\to j} + b e_j)\| \\
&\qquad\times \left|\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta) \exp\left(-\frac{(b - b_{\theta,j}(\boldsymbol{\beta}))^2 \|e_j\|^2_\theta}{2}\right) \frac{\|e_j\|_\theta}{\sqrt{2\pi}}\right)\right| \mathrm{d}b.
\end{aligned} \tag{30}$$

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

$$\begin{aligned}
&\left|\frac{\mathrm{d}}{\mathrm{d}\epsilon} \exp\left(-\frac{(b - b_{\theta,j}(\boldsymbol{\beta}))^2 \|e_j\|^2_\theta}{2}\right) \frac{\|e_j\|_\theta}{\sqrt{2\pi}}\right| \\
&\quad\le C_{\mathcal{K}} \left(1 + |b - b_{\theta,j}(\boldsymbol{\beta})|\right)^2 \exp\left(-\frac{(b - b_{\theta,j}(\boldsymbol{\beta}))^2 \|e_j\|^2_\theta}{2}\right) \\
&\qquad\times \frac{\|e_j\|_\theta}{\sqrt{2\pi}} \left(\left|\frac{\mathrm{d}}{\mathrm{d}\epsilon} b_{\theta,j}(\boldsymbol{\beta})\right| + \left|\frac{\mathrm{d}}{\mathrm{d}\epsilon} \|e_j\|_\theta\right|\right).
\end{aligned} \tag{31}$$

Since $\frac{\mathrm{d}}{\mathrm{d}\epsilon}\|e_j\|_\theta = \frac{1}{2\|e_j\|_\theta} e_j^t \frac{\mathrm{d}}{\mathrm{d}\epsilon} \Gamma_\theta^{-1} e_j$, $\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Gamma_\theta^{-1} = -\Gamma_\theta^{-1} \frac{\mathrm{d}}{\mathrm{d}\epsilon}\Gamma_\theta \Gamma_\theta^{-1}$ and $\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Gamma_\theta = \frac{s'_3 - s_3}{n + a_g}$ (see (3)), we deduce that there exists another constant $C_{\mathcal{K}}$ such that

$$\left|\frac{\mathrm{d}}{\mathrm{d}\epsilon}\|e_j\|_\theta\right| \le C_{\mathcal{K}} \|s' - s\|. \tag{32}$$

Similarly, updating the constant $C_{\mathcal{K}}$, we have[1]

$$\left|\frac{\mathrm{d}}{\mathrm{d}\epsilon} b_{\theta,j}(\boldsymbol{\beta})\right| \leq C_{\mathcal{K}}(1+\|\boldsymbol{\beta}\|)\|s'-s\|. \tag{33}$$

Now, concerning the derivative of $\tilde{r}_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)$, since

$$\log(\tilde{r}_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)) = \frac{1}{2}\sum_{i=1}^{n}(\|y_i - K_p^{\tilde{\boldsymbol{\beta}}_i}\alpha\|^2 - \|y_i - K_p^{\boldsymbol{\beta}_i}\alpha\|^2)$$

with $\tilde{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_{i,b\to j}$, i corresponding to the ith image, only one term of the previous sum is non-zero. We deduce from the fact that K_p is bounded and from (3) that $|\frac{\mathrm{d}}{\mathrm{d}\epsilon}\log(\tilde{r}_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta))| \leq C_{\mathcal{K}}|\frac{\mathrm{d}}{\mathrm{d}\epsilon}\alpha| \leq C_{\mathcal{K}}\|s-s'\|$, so that, using the facts that $\tilde{r}_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)$ is uniformly bounded for $\theta \in \hat{\theta}(\mathcal{K})$, $\boldsymbol{\beta} \in \mathbb{R}^N$ and $b \in \mathbb{R}$, there exists a new constant $C_{\mathcal{K}}$ such that

$$\left|\frac{\mathrm{d}}{\mathrm{d}\epsilon}\tilde{r}_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)\right| \leq C_{\mathcal{K}}\|s-s'\|.$$

Thus, using (31), (32) and (33), we get, for a new constant $C_{\mathcal{K}}$, that

$$\begin{aligned}
&\left|\frac{\mathrm{d}}{\mathrm{d}\epsilon} r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)\exp\left(-\frac{(b-b_{\theta,j}(\boldsymbol{\beta}))^2\|e_j\|_\theta^2}{2}\right)\frac{\|e_j\|_\theta}{\sqrt{2\pi}}\right| \\
&\quad\leq C_{\mathcal{K}}(1+\|\boldsymbol{\beta}\|)\|s'-s\|\big(1+|b-b_{\theta,j}(\boldsymbol{\beta})|\big)^2\exp\left(-\frac{(b-b_{\theta,j}(\boldsymbol{\beta}))^2\|e_j\|_\theta^2}{2}\right)\frac{\|e_j\|_\theta}{\sqrt{2\pi}}.
\end{aligned}$$

Since $\|f(\boldsymbol{\beta})\| \leq \|f\|_{V^p} V^p(\boldsymbol{\beta})$ and $V(a+b) = 1 + \|a+b\|^2 \leq 2(V(a)+V(b))$, we have $\|f(\boldsymbol{\beta}_{0\to j} + be_j)\| \leq C\|f\|_{V^p}(V^p(\boldsymbol{\beta}_{0\to j}) + V^p(be_j))$ with $C = 2^{2p-1}$. Hence, there exists a $C_{\mathcal{K}}$ such that $\forall (s,s') \in \mathcal{K}^2$, $\forall 1 \leq j \leq N$, $\forall \boldsymbol{\beta} \in \mathbb{R}^N$ and $\forall \epsilon \in [0,1]$,

$$\begin{aligned}
&\int_{\mathbb{R}} \|f(\boldsymbol{\beta}_{0\to j} + be_j)\|\left|\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(r_j(\boldsymbol{\beta}^j, b; \boldsymbol{\beta}^{-j}, \theta)\exp\left(-\frac{(b-b_{\theta,j}(\boldsymbol{\beta}))^2\|e_j\|_\theta^2}{2}\right)\frac{\|e_j\|_\theta}{\sqrt{2\pi}}\right)\right|\mathrm{d}b \\
&\quad\leq C_{\mathcal{K}}\|f\|_{V^p} V^p(\boldsymbol{\beta}_{0\to j})(1+\|\boldsymbol{\beta}\|)\|s'-s\| \leq C_{\mathcal{K}}\|f\|_{V^p} V^p(\boldsymbol{\beta})(1+\|\boldsymbol{\beta}\|)\|s'-s\|,
\end{aligned}$$

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

$$\left\|\left(\boldsymbol{\Pi}^2_{\theta(1),j} - \boldsymbol{\Pi}^2_{\theta(0),j}\right)f(\boldsymbol{\beta})\right\| \leq C_{\mathcal{K}}\|f\|_{V^p} V^{p+1/2}(\boldsymbol{\beta})\|s'-s\|. \tag{34}$$

[1]Note that the extra factor $(1+\|\boldsymbol{\beta}\|)$ appearing in the right-hand side of (33), compared to the right-hand side of (32), alleviates the need to show the usual Lipschitz condition $\|\boldsymbol{\Pi}_{\hat{\theta}(s)} f - \boldsymbol{\Pi}_{\hat{\theta}(s')} f\|_{V^{p'}} \leq C_{\mathcal{K}}\|f\|_{V^q}\|s-s'\|$ with $q = p$. Weaker Lipschitz conditions, such as condition A3′(ii) of Theorem 1 are needed.

Now, looking at the first term in (6.3), we easily deduce from the previous study for $f \equiv f(\boldsymbol{\beta})$ that

$$\begin{aligned}\left\|\left(\boldsymbol{\Pi}^1_{\theta(1),j} - \boldsymbol{\Pi}^1_{\theta(0),j}\right) f(\boldsymbol{\beta})\right\| &\leq C_{\mathcal{K}} V(\boldsymbol{\beta})^{1/2} \|s' - s\| \|f(\boldsymbol{\beta})\| \\ &\leq C_{\mathcal{K}} \|f\|_{V^p} V^{p+1/2}(\boldsymbol{\beta}) \|s' - s\|\end{aligned} \tag{35}$$

so that adding (34) and (35), we get (again updating $C_{\mathcal{K}}$) that

$$\left\|\left(\boldsymbol{\Pi}_{\theta(1),j} - \boldsymbol{\Pi}_{\theta(0),j}\right) f\right\|_{V^{p+1/2}} \leq C_{\mathcal{K}} \|f\|_{V^p} \|s' - s\|. \tag{36}$$

We conclude the proof by stating that $\boldsymbol{\Pi}_{\theta(1)} - \boldsymbol{\Pi}_{\theta(0)} = \sum_{j=1}^N \boldsymbol{\Pi}_{\theta(1),j+1,N} \circ (\boldsymbol{\Pi}_{\theta(1),j} - \boldsymbol{\Pi}_{\theta(0),j}) \circ \boldsymbol{\Pi}_{\theta(0),1,j-1}$, where $\boldsymbol{\Pi}_{\theta,q,r} = \boldsymbol{\Pi}_{\theta,r} \circ \boldsymbol{\Pi}_{\theta,r-1} \circ \cdots \circ \boldsymbol{\Pi}_{\theta,q}$ for any integer $q \leq r$ and any $\theta \in \Theta$ so that, using (6.2) and (36), the result is straightforward. □

Lemma 5. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. There exists a constant $C_{\mathcal{K}}$ such that for all $p \geq 1$ and any function $f \in \mathcal{L}_{V^p}$, $\forall (s, s') \in \mathcal{K}^2$, $\forall k \geq 0$, we have, for $\theta = \hat{\theta}(s)$ and $\theta' = \hat{\theta}(s')$, that*

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

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

$$\boldsymbol{\Pi}^k_\theta f - \boldsymbol{\Pi}^k_{\theta'} f = \sum_{i=1}^{k-1} \boldsymbol{\Pi}^i_\theta (\boldsymbol{\Pi}_\theta - \boldsymbol{\Pi}_{\theta'}) \left(\boldsymbol{\Pi}^{k-i-1}_{\theta'} f - \boldsymbol{\pi}_{\theta'}(f)\right).$$

Using Lemma 4, the fact that $\|\boldsymbol{\Pi}^k_\theta (f - \pi_\theta(f))\|_{V^p} \leq \gamma^k_{\mathcal{K}} \|f\|_{V^p}$ with $\gamma_{\mathcal{K}} < 1$ (geometric ergodicity) and $\sup_{j \geq 0} \sup_{\theta \in K} \|\boldsymbol{\Pi}^j_\theta V^q\|_{V^q} < \infty$, we get

$$\begin{aligned}\|\boldsymbol{\Pi}^k_\theta f - \boldsymbol{\Pi}^k_{\theta'} f\|_{V^{p+1/2}} &\leq C_{\mathcal{K}} \sum_{i=1}^{k-1} \left\|(\boldsymbol{\Pi}_\theta - \boldsymbol{\Pi}_{\theta'}) \left(\boldsymbol{\Pi}^{k-i-1}_{\theta'} f - \boldsymbol{\pi}_{\theta'}(f)\right)\right\|_{V^{p+1/2}} \\ &\leq C_{\mathcal{K}} \|f\|_{V^p} |s - s'| \sum_{i=1}^{k-1} \gamma^{k-i+1}_{\mathcal{K}}\end{aligned}$$

and the lemma is proved. □

We now prove that h is a Hölder function, linearly adapting Appendix B of Andrieu, Moulines and Priouret (2005).

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

$$\begin{aligned}A(s, s') &= \left(h(s) - \boldsymbol{\Pi}^k_\theta H_s(\boldsymbol{\beta})\right) + \left(\boldsymbol{\Pi}^k_{\theta'} H_{s'}(\boldsymbol{\beta}) - h(s')\right), \\ B(s, s') &= \boldsymbol{\Pi}^k_\theta H_s(\boldsymbol{\beta}) - \boldsymbol{\Pi}^k_{\theta'} H_s(\boldsymbol{\beta}), \\ C(s, s') &= \boldsymbol{\Pi}^k_{\theta'} H_s(\boldsymbol{\beta}) - \boldsymbol{\Pi}^k_{\theta'} H_{s'}(\boldsymbol{\beta}).\end{aligned}$$

Using geometric ergodicity, Lemmas 4 and 5, we get that there exists an $C > 0$, independent of k, such that

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

This yields

$$\|h(s)-h(s')\| \le CV^{3/2}(\boldsymbol{\beta})(\gamma^k + \|s-s'\|).$$

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

We can now complete the proof of A3′(ii). On one hand, we have

$$\begin{aligned}&\left\|\left(\boldsymbol{\Pi}_\theta^k H_s(\boldsymbol{\beta}) - h(s)\right) - \left(\boldsymbol{\Pi}_{\theta'}^k H_{s'}(\boldsymbol{\beta}) - h(s')\right)\right\| \\ &\quad\le \|\boldsymbol{\Pi}_\theta^k H_s(\boldsymbol{\beta}) - \boldsymbol{\Pi}_\theta^k H_{s'}(\boldsymbol{\beta})\| \\ &\qquad + \|\boldsymbol{\Pi}_\theta^k H_{s'}(\boldsymbol{\beta}) - \boldsymbol{\Pi}_{\theta'}^k H_{s'}(\boldsymbol{\beta})\| + \|h(s)-h(s')\| \le C\|s-s'\| V^{3/2}(\boldsymbol{\beta}).\end{aligned}$$

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

$$\left\|\left(\boldsymbol{\Pi}_\theta^k H_s(\boldsymbol{\beta}) - h(s)\right) - \left(\boldsymbol{\Pi}_{\theta'}^k H_{s'}(\boldsymbol{\beta}) - h(s')\right)\right\| \le C\gamma^k V^{3/2}(\boldsymbol{\beta}).$$

Hence, for any $t \ge 0$ and $T \ge t$, we have

$$\begin{aligned}\left\|\boldsymbol{\Pi}_\theta^t g_{\hat\theta(s)}(\boldsymbol{\beta}) - \boldsymbol{\Pi}_{\theta'}^t g_{\hat\theta(s')}(\boldsymbol{\beta})\right\| &\le \sum_{k=t}^{\infty} \left\|\left(\boldsymbol{\Pi}_\theta^k H_s(\boldsymbol{\beta}) - h(s)\right) - \left(\boldsymbol{\Pi}_{\theta'}^k H_{s'}(\boldsymbol{\beta}) - h(s')\right)\right\| \\ &\le CV^{3/2}(\boldsymbol{\beta})\left[T\|s-s'\| + \frac{\gamma^{T+t}}{1-\gamma}\right].\end{aligned}$$

Setting $T = [\log\|s-s'\|/\log(\gamma)]$ for $\|s-s'\| \le \delta < 1$ and $T = t$ otherwise, also using 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 $a < 1$.

We finally focus on the proof of A3′(iii). Once again, we first prove a specific result for each function V_θ and then obtain a result for the function V.

Lemma 6. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$ and $p \ge 1$. There exists $C_{\mathcal{K},p} > 0$ such that for any s, $s' \in \mathcal{K}$, for any $\boldsymbol{\beta} \in \mathbb{R}^N$,*

$$\left|V_{\hat\theta(s)}^p(\boldsymbol{\beta}) - V_{\hat\theta(s')}^p(\boldsymbol{\beta})\right| \le C_{\mathcal{K},p}\|s-s'\| V_{\hat\theta(s)}^p(\boldsymbol{\beta}).$$

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 a C such that $\forall (s, s') \in \mathcal{K}^2$, $|\Gamma_g^{-1} - (\Gamma'_g)^{-1}| \leq C\|s - s'\|$ and

$$\left|V_{\hat{\theta}(s)}(\boldsymbol{\beta}) - V_{\hat{\theta}(s')}(\boldsymbol{\beta})\right| \leq \sum_{i=1}^{n} \beta_i^t \left(\Gamma_g^{-1} - (\Gamma'_g)^{-1}\right)\beta_i \leq C\|s - s'\| V(\boldsymbol{\beta}).$$

The result follows from the existence of a constant C such that $\frac{1}{c} V(\boldsymbol{\beta}) \leq V_{\hat{\theta}(s)}(\boldsymbol{\beta}) \leq C V(\boldsymbol{\beta})$ for any $(\boldsymbol{\beta}, s) \in \mathbb{R}^N \times \mathcal{K}$. □

Lemma 7. *Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$ and $p \geq 1$. There exist $\bar{\varepsilon} > 0$ and $C > 0$ such that for any sequence $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k\geq 0}$ such that $\varepsilon_k \leq \bar{\varepsilon}$ for k large enough, any sequence $\boldsymbol{\Delta} = (\Delta_k)_{k\geq 0}$ and any $\boldsymbol{\beta} \in \mathbb{R}^N$,*

$$\sup_{s\in\mathcal{K}} \sup_{k\geq 0} \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[V^p(\boldsymbol{\beta}_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}\right] \leq C V^p(\boldsymbol{\beta}).$$

Proof. Let K be a compact subset of Θ such that $\hat{\theta}(\mathcal{K}) \subset K$. We note, in the sequel, that $\theta_k = \hat{\theta}(s_k)$. For $k \geq 2$, we have, using the Markov property and Lemmas 2 and 6,

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[V^p_{\theta_{k-1}}(\boldsymbol{\beta}_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}\right] \\
&\quad\leq \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[\boldsymbol{\Pi}_{\theta_{k-1}} V^p_{\theta_{k-1}}(\boldsymbol{\beta}_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}\right] \\
&\quad\leq \rho\left(\mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[V^p_{\theta_{k-2}}(\boldsymbol{\beta}_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}\right] + \mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[\left(V^p_{\theta_{k-1}}(\boldsymbol{\beta}_{k-1}) - V^p_{\theta_{k-2}}(\boldsymbol{\beta}_{k-1})\right)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}\right]\right) \\
&\qquad + C \\
&\quad\leq \rho\left(\mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[V^p_{\theta_{k-2}}(\boldsymbol{\beta}_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k-1}\right] + C'\epsilon_{k-1}\mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[V^p_{\theta_{k-2}}(\boldsymbol{\beta}_{k-1})\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k-1}\right]\right) + C.
\end{aligned}$$

By induction, we show that

$$\mathbb{E}^{\boldsymbol{\Delta}}_{\boldsymbol{\beta},s}\left[V^p_{\theta_{k-1}}(\boldsymbol{\beta}_k)\mathbb{1}_{\sigma(\mathcal{K})\wedge\nu(\boldsymbol{\varepsilon})\geq k}\right] \leq \prod_{l=1}^{k-1}\left(\rho(1 + C'\varepsilon_l)\right) V^p_{\hat{\theta}(s)}(\boldsymbol{\beta}) + \frac{C}{(1 - \rho(1 + C'\bar{\varepsilon}))}.$$

Choosing $\bar{\varepsilon}$ such that $\rho(1 + C'\bar{\varepsilon}) < 1$ and again introducing $0 \leq c_1 \leq c_2$ such that $c_1 V(\boldsymbol{\beta}) \leq V_\theta(\boldsymbol{\beta}) \leq c_2 V(\boldsymbol{\beta})$ for any $(\boldsymbol{\beta}, \theta) \in \mathbb{R}^N \times K$ completes the proof. □

This yields A3′(iii).
This concludes the demonstration of Theorem 2.

7. Conclusion and discussion

We have proposed a stochastic algorithm for constructing Bayesian non-rigid deformable models in the same context as Allassonnière, Amit and Trouvé (2007), together with a proof of convergence toward a critical point of the observed likelihood. To the best of our knowledge, this is

the first theoretical result on convergence in the context of deformable templates. The algorithm is based on a stochastic approximation of the EM algorithm using an MCMC approximation of the posterior distribution and truncation on random boundaries. Although our main contribution is theoretical, the preliminary experiments presented here on the United States Postal Service database show that the stochastic approach can be easily implemented and is robust to noisy situations, yielding better results than the previous deterministic schemes.

Many interesting questions remain open. One may ask, what is the convergence rate of such stochastic algorithms? A first result has been proven in Delyon, Lavielle and Moulines (1999) for the standard SAEM algorithm. Under mild conditions, the authors state a central limit theorem for an average sequence of the estimated parameters $(\theta_k)_k$. Concerning the generalization when introducing MCMC, a first step has been tackled in Andrieu and Moulines (2006). Under some restrictive assumptions, the authors can prove a central limit theorem for an ergodic adaptive Monte Carlo Markov chain. We believe that it is possible to obtain these kinds of convergence rates for the SAEM-MCMC algorithm proposed in this paper.

Another question refers to the extension of the stochastic scheme to mixtures of deformable models (defined as the multicomponent model in Allassonnière, Amit and Trouvé (2007)), 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 for real data analysis where the restriction to a unique deformable model could be too limiting. The design of such mixtures 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 would appear at first glance: due to the high-dimensional hidden deformation variables, a naive extension of the Markovian dynamics to the component variables will have extremely poor mixing properties, leading to an impractical algorithm. A less straightforward extension involving multiple MCMC chains is currently being studied.

Another interesting extension is to consider diffeomorphic mappings 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, in principle, be done using tangent models based on geodesic shooting, in the spirit of Vaillant et al. (2004).

References

Allassonière, S., Amit, Y., Kuhn, E. and Trouvé, A. (2006). Generative model and consistent estimation algorithms for non-rigid deformable models. In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing* **5**. IEEE.

Allassonnière, S., Amit, Y. and Trouvé, A. (2007). Toward a coherent statistical framework for dense deformable template estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 3–29. MR2301497

Amit, Y. (1996). Convergence properties of the Gibbs sampler for perturbations of Gaussians. *Ann. Statist.* **24** 122–140. MR1389883

Amit, Y., Grenander, U. and Piccioni, M. (1991). Structural image restoration through deformable template. *J. Amer. Statist. Assoc.* **86** 376–387.

Andrieu, C. and Moulines, É. (2006). On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.* **16** 1462–1505. MR2260070

Andrieu, C., Moulines, É. and Priouret, P. (2005). Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.* **44** 283–312 (electronic). MR2177157

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

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** 94–128. MR1701103

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **1** 1–22. MR0501537

Douc, R., Moulines, E. and Rosenthal, J.S. (2004). Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Probab.* **14** 1643–1665. MR2099647

Glasbey, C.A. and Mardia, K.V. (2001). A penalised likelihood approach to image warping. *J. Roy. Statist. Soc. Ser. B* **63** 465–492. MR1858399

Grenander, U. and Miller, M.I. (1998). Computational anatomy: An emerging discipline. *Quart. Appl. Math.* **LVI** 617–694. MR1668732

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM Probab. Stat.* **8** 115–131 (electronic). MR2085610

Marsland, S., Twining, C. and Taylor, C. (2007). A minimum description length objective function for groupwise non rigid image registration. *Image and Vision Computing* **26** 333–346.

Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. *Communications and Control Engineering Series*. London: Springer. MR1287609

Richard, F., Samson, A. and Cuénod, C. (2009). A saem algorithm for the estimation of template and deformation parameters in medical image sequences. *Statist. Comput.* **19** 465–478.

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

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

Received June 2007 and revised December 2008