Component-wise Approximate Bayesian Computation via Gibbs-like steps

BY GRÉGOIRE CLARTÉ, CHRISTIAN P. ROBERT, ROBIN J. RYDER, AND JULIEN STOEHR

Université Paris-Dauphine, Université PSL, CNRS, CEREMADE, 75016 Paris, France

SUMMARY

Approximate Bayesian computation methods are useful for generative models with intractable likelihoods. These methods are however sensitive to the dimension of the parameter space, requiring exponentially increasing resources as this dimension grows. To tackle this difficulty, we explore a Gibbs version of the Approximate Bayesian computation approach that runs component-wise approximate Bayesian computation steps aimed at the corresponding conditional posterior distributions, and based on summary statistics of reduced dimensions. While lacking the standard justifications for the Gibbs sampler, the resulting Markov chain is shown to converge in distribution under some partial independence conditions. The associated stationary distribution can further be shown to be close to the true posterior distribution and some hierarchical versions of the proposed mechanism enjoy a closed form limiting distribution. Experiments also demonstrate the gain in efficiency brought by the Gibbs version over the standard solution.

Some key words: Approximate Bayesian computation, Gibbs sampler, hierarchical Bayes model, curse of dimensionality, conditional distributions, convergence of Markov chains, incompatible conditionals.

1. INTRODUCTION

Approximation Bayesian computation (ABC) is a computational method which stemmed from population genetics to deal with intractable likelihoods, that is model with intractable likelihood but which can be simulated from (Tavaré et al., 1997; Beaumont et al., 2002). It has been since then applied to numerous other fields: see for example Toni et al. (2008); Csilléry et al. (2010); Moores et al. (2015); Sisson et al. (2018). The principle of the method is to simulate pairs of parameters and pseudo-data from the prior predictive, keeping only the parameters that bring the pseudo-data close enough to the observed data. Proximity is often defined in terms of a projection of the data, called a summary statistic. From the start, this method has suffered from the curse of dimensionality in that the dimension of the parameter to be inferred imposes a lower bound on the dimension of the corresponding summary statistic to be used (results by Fearnhead & Prangle (2012) and Li & Fearnhead (2018) imply that the dimension of the summary should be identical to the dimension of the parameter). This constraint impacts the range of the distance between observed and simulated summaries, with the distance choice having a growing impact as the dimension increases. Reducing the dimension of the summary is thus impossible without reducing the dimension of the parameter, which sounds an impossible goal unless one infers about one parameter at a time, suggesting a Gibbs sampling strategy where a different and much reduced dimension summary statistic is used for each component of the parameter. The purpose of this paper is to explore and validate this strategy.

Additionally, the Gibbs perspective allows to account for the current values of the other components of the parameter and therefore to shy away from simulating from the prior which is an inefficient proposal. This feature connects this proposal with earlier solutions in the literature such as the Metropolis version of Marjoram et al. (2003) and the various sequential Monte Carlo schemes (Toni et al., 2008; Beaumont et al., 2009).

A Gibbs version of the ABC method offers a range of potential improvements compared with earlier versions, induced in most cases by the dimension reduction thus achieved. First, in hierarchical Bayes
models, conditioning shortens the number of dependent components, and some of the conditionals may be available in closed form, which makes the approach only semi-approximate. Second, since the conditional targets live in small dimension spaces, they can more easily be parametrised by low dimension functions of the conditioning terms. This justifies using a restricted range of collection of statistics, which may in addition depend on other parameters. Third, reducing the dimension of the summary statistic produces an improvement in the approximation since a smaller tolerance can then be handled at a manageable computing cost.

In Section 2, we define the ensuing algorithm, called ABC-Gibbs, and validate it in the general framework of a generic multi-parameter model. Sections 3 and 5 focus on theoretical and numerical results for two types of hierarchical models. Such models appear naturally in intractable inference and allow to restrict the call to approximate conditionals to some parameters. Section 4 presents numerical simulations for a more complex model, used to described the daily flux emitted by some stellar objects using a moving average model. Finally, in Section 6, we tackle the issue of the incompatibility between the resulting conditionals and produce a formal alternative to the working version of the algorithm, towards highlighting the difference with the original ABC method, as well as suggesting a post-processing correction.

2. APPROXIMATE BAYESIAN GIBBS SAMPLING

2.1. Vanilla approximate Bayesian computation

Approximate Bayesian computation methods, summarised in Algorithm 1, provide a technique to sample posterior distributions when the corresponding likelihood $f(x|\theta)$ is intractable, that is the numerical value $f(x|\theta)$ cannot be computed in a reasonable amount of time, but the model is generative, that is it allows for the generation of synthetic data given a value of the parameter. Given a prior distribution on the parameter $\theta$, it builds upon samples from the associated prior predictive $(\theta^{(i)}, x^{(i)})_{i=1,...,N}$ by selecting pairs such that the pseudo-data $x^{(i)}$ stands in a neighbourhood of the observed data $x^\star$.

Since both the simulated and observed dataset may belong to a space of a high dimension, the neighbourhood is usually defined with respect to a summary statistic $s(\cdot)$ of a lesser dimension and an associated distance $d$ (see Marin et al. (2012) for a review).

**Algorithm 1. Vanilla Approximate Bayesian computation**

**Input:** observed dataset $x^\star$, number of iterations $N$, threshold $\varepsilon > 0$, summary statistic $s$.

**Output:** a sample $(\theta^{(1)}, \ldots, \theta^{(N)})$.

for $i = 1, \ldots, N$ do

\[ \text{repeat} \]

\[ \theta^{(i)} \sim \pi(\cdot) \]

\[ x^{(i)} \sim f(\cdot | \theta^{(i)}) \]

\[ \text{until } d(s(x^{(i)}), s(x^\star)) < \varepsilon \]

The output of Algorithm 1 is a sample distributed from an approximation of the posterior (Tavaré et al., 1997; Sisson et al., 2018). Its density can be written as

$$\pi_\varepsilon(\theta | x^\star, s) \propto \int \pi(\theta)f(x | \theta)1_{d(s(x), s(x^\star))<\varepsilon} \, dx$$

This approximation depends on the choice of both the summary statistic $s$ and the tolerance level $\varepsilon$. Frazier et al. (2018) show its consistency, namely that when the number of observations tends to $\infty$ and the tolerance tends to 0 at a proper relative rate, the approximate posterior concentrates at the true value of the parameter, albeit as a posterior distribution associated with the statistic $s$, rather than the true posterior, when $s$ is not sufficient. The shape of the asymptotic distribution is further discussed in Li & Fearnhead (2018) and Frazier et al. (2018).

More to the point, given a fixed number of observations, the approximate posterior also converges to the true posterior $\pi(\theta | s(x^\star))$, by opposition to the classical posterior $\pi(\theta | x^\star)$, when the tolerance level
goes to 0. In practice, however, the tolerance level cannot be equal to zero and is customarily chosen as a simulated distance quantile. That is, once a prior predictive sample, also called a reference table, is produced, the tolerance is derived as an empirical quantile from the resulting distance between summary statistics of the simulated and observed data.

2.2. Gibbs sampler

The Gibbs sampler, first introduced by Geman & Geman (1984) and generalised by Gelfand & Smith (1990), is an essential element in Markov chain Monte Carlo methods (Robert & Casella, 2004; Gelman et al., 2013). As described in Algorithm 2, for a parameter \( \theta = (\theta_1, \ldots, \theta_n) \), it produces a Markov chain associated with a given target joint distribution, denoted \( \pi \), by alternatively sampling from each of its conditionals.

Algorithm 2. Gibbs sampler

**Input:** number of iterations \( N \), starting point \( \theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)}) \).

**Output:** a sample \((\theta^{(1)}(\cdot), \ldots, \theta^{(N)}(\cdot))\).

for \( i = 1, \ldots, N \) do

for \( j = 1, \ldots, n \) do

\[ \theta_j^{(i)} \sim \pi(\cdot | \theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \ldots, \theta_n^{(i-1)}) \]

Gibbs sampling is well suited to high-dimensional situations where the conditional distributions are easy to sample. In particular, as illustrated by the long-lasting success of the BUGS software (Lunn et al., 2010), hierarchical Bayes models often allow for simplified conditional distributions thanks to partial independence properties. Considering for instance the common hierarchical model (Lindley & Smith, 1972; Carlin & Louis, 1996) defined by

\[
x_j | \mu_j \sim \pi(x_j | \mu_j), \quad \mu_j \sim \pi(\mu_j | \alpha), \quad \alpha \sim \pi(\alpha).
\]

The joint posterior of \( \mu = (\mu_1, \ldots, \mu_n) \) conditional on \( \alpha \) then factorises as

\[
\pi(\mu | x_1, \ldots, x_n, \alpha) = \prod_{j=1}^n \pi(\mu_j | \alpha) \pi(x_j | \mu_j).
\]

This implies that the full conditional posteriors of each \( \mu_j \) only depends on \( \alpha \) and \( x_j \), independently of the other \((\mu_\ell, x_\ell)\)'s.

2.3. Component-wise Approximate Bayesian Computation

When handling a model such as (1) with both a high-dimensional parameter and an intractable likelihood, the Gibbs sampler cannot be implemented, while the vanilla ABC sampler is highly inefficient. This curse of dimensionality attached to the ABC algorithm is well documented (Li & Fearnhead, 2018).

Bringing both approaches together may subdue this loss efficiency, by sequentially sampling from the ABC version of the conditionals. Each step in Algorithm 2 is then replaced by a call to Algorithm 1, conditional on the other components of the parameter. We obtain a generic componentwise approximate Bayesian computational method, summarised as Algorithm 3. This algorithm can be analysed as a variation of Algorithm 1 in which the synthetic data \( x^{(i)} \) are simulated from the conditional posterior predictive, rather than from the prior predictive. This may result in simulating both parameters and pseudo-data component-wise from spaces of smaller dimensions. This also allows the use of statistics of lower dimensions, as exemplified in Section 5.

If \( \varepsilon_j = 0 \) and if \( s_j \) is a conditionally sufficient statistic, the corresponding \( j \)th step in Algorithm 3 is an exact simulation from the corresponding posterior. Thus, if some of the conditional distributions are generative models, they remove the need for an approximate step in the algorithm. In practice, to simulate from the approximate conditional, and similarly to Algorithm 1, we take \( \varepsilon_j \) as an empirical distance quantile. In other words, for the \( j \)th component of the parameter, conditional on the other components, we
Algorithm 3. ABC-Gibbs

\textbf{Input:} number of iterations $N$, starting point $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$, threshold $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, statistics $s_1, \ldots, s_n$.

\textbf{Output:} a sample $(\theta^{(1)}, \ldots, \theta^{(N)})$

\begin{algorithmic}
\FOR{$i = 1, \ldots, N$}
  \FOR{$j = 1, \ldots, n$}
    $\theta_j^{(i)} \sim \pi_{\varepsilon_j}(\cdot | x^*, s_j, \theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta_{j+1}^{(i-1)}, \ldots, \theta_n^{(i-1)})$
  \ENDFOR
\ENDFOR
\end{algorithmic}

simulate a small reference table from its conditional prior and output the parameter associated with the smallest distance.

At first, the purpose of this algorithm may sound unclear as the limiting distribution and its existence are unknown. Both issues are discussed below with further emphasis in Section 6 on the possible lack of compatibility of the approximate conditionals and the limiting distribution associated with an alternative version. Convergence can indeed be achieved, based on a simple condition, with a proof provided in the Supplementary Material, Section 7. For simplicity’s sake, we only consider the case when $n = 2$ in Algorithm 3.

\textbf{Theorem 1.} Assume that there exists $0 < \kappa < 1/2$ such that

$$\sup_{\theta_1, \hat{\theta}_1} \| \pi_{\varepsilon_2}(\cdot | x^*, s_2, \theta_1) - \pi_{\varepsilon_2}(\cdot | x^*, s_2, \hat{\theta}_1) \|_{TV} = \kappa.$$\n
The Markov chain produced by Algorithm 3 then converges geometrically in total variation distance to a stationary distribution $\nu_\varepsilon$, with geometric rate $1 - 2\kappa$.

The above assumption is verified in particular when the parameter space is compact. Possible relaxations are not covered in this paper. The corresponding assumption for the general case $n > 2$ is given in the Supplementary Material, Section 7.

The limiting distribution $\nu_\varepsilon$ is not necessarily a standard posterior. We can however provide an evaluation of the distance between $\nu_\varepsilon$ and the limiting distribution $\nu_0$ of Algorithm 3 with $\varepsilon_1 = \varepsilon_2 = 0$. In a compact parameter space, $\nu_0$ always exists, but it may differ from the joint distribution associated with a vanilla Gibbs sampler, because the conditionals may be based on different summary statistics $s_1$ and $s_2$. Section 6 explores the meaning of $\nu_0$ by considering an alternative ABC-Gibbs algorithm that allows for a corrective post-processing.

\textbf{Theorem 2.} Assume that

$$L_0 = \sup_{\varepsilon_2, \theta_1, \hat{\theta}_1} \sup_{\varepsilon_2} \| \pi_{\varepsilon_2}(\cdot | x^*, s_2, \theta_1) - \pi_0(\cdot | x^*, s_2, \hat{\theta}_1) \|_{TV} < 1/2,$$

$$L_1(\varepsilon_1) = \sup_{\theta_2} \| \pi_{\varepsilon_1}(\cdot | x^*, s_1, \theta_2) - \pi_0(\cdot | x^*, s_1, \theta_2) \|_{TV} \xrightarrow{\varepsilon_1 \rightarrow 0} 0,$$

$$L_2(\varepsilon_2) = \sup_{\hat{\theta}_1} \| \pi_{\varepsilon_2}(\cdot | x^*, s_2, \theta_1) - \pi_0(\cdot | x^*, s_2, \theta_1) \|_{TV} \xrightarrow{\varepsilon_2 \rightarrow 0} 0.$$

Then

$$\| \nu_\varepsilon - \nu_0 \|_{TV} \leq \frac{L_1(\varepsilon_1) + L_2(\varepsilon_2)}{1 - 2L_0} \xrightarrow{\varepsilon \rightarrow 0} 0.$$\n
The assumptions found in both Theorem 1 and 2 may prove delicate to interpret or to verify in realistic situations. The next sections reformulate these results for two common situations.
3. Component-wise approximate Bayesian computation: the hierarchical case

3.1. Algorithm and theory

In this section, we focus on the two-stage simple hierarchical Bayes model given in (1), where the conditional distributions greatly simplify as $\pi(\mu_j \mid x^*, \alpha, \mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_n) = \pi(\mu_j \mid x^*_j, \alpha)$ and $\pi(\alpha \mid \mu, x^*) = \pi(\alpha \mid \mu)$. In this particular case Algorithm 3 writes as Algorithm 4. (To simulate from the approximate conditional distributions, we might resort to a Metropolis step, using the prior distribution as proposal.)

Algorithm 4. ABC-Gibbs sampler for hierarchical model (1)

**Input:** observed dataset $x^*$, number of iterations $N$, starting points $\alpha^{(0)}$ and $\mu^{(0)} = (\mu_1^{(0)}, \ldots, \mu_n^{(0)})$, thresholds $\varepsilon_\alpha$ and $\varepsilon_\mu$, summary statistics $s_\alpha$ and $s_\mu$.

**Output:** A sample $((\alpha^{(i)}, \mu^{(i)}))_{1 \leq i \leq N}$.

for $i = 1, \ldots, N$ do

for $j = 1, \ldots, n$ do

$\mu_j^{(i)} \sim \pi_{\epsilon_\mu}(\cdot \mid x^*_j, s_\mu, \alpha^{(i-1)})$

$\alpha^{(i)} \sim \pi_{\epsilon_\alpha}(\cdot \mid \mu^{(i)}, s_\alpha)$

Since the parameter $\mu$ in model (1) does not necessarily live in a compact set, it may prove difficult to apply Theorem 1 as is. We therefore rewrite Theorem 1 with simpler hypotheses attuned to model (1). For the sake of simplicity, we assume that $n = 1$ in what follows; the general case $n > 1$ is once again a simple extension.

**Theorem 3.** Assume there exists a non-empty convex set $C$ with positive prior measure such that

$$\begin{align*}
\kappa_1 &= \inf_{s_\alpha(\mu) \in C} \pi(B_{s_\alpha(\mu), \varepsilon_\mu/4}) > 0, \\
\kappa_2 &= \inf_{\alpha \in s_\alpha(\mu) \in C} \pi_{\epsilon_\mu}(B_{s_\alpha(\mu), 3\varepsilon_\mu/2} \mid x^*, s_\mu, \alpha) > 0, \\
\kappa_3 &= \inf_{\alpha \in \pi_{\epsilon_\mu}(s_\mu(\mu) \in C \mid x^*, s_\mu, \alpha) > 0,}
\end{align*}$$

where $B_{z,h}$ denotes the ball of center $z$ and radius $h$. Then the Markov chain produced by Algorithm 4 converges geometrically in total variation distance to a stationary distribution $\nu_\nu$, with geometric rate $1 - \kappa_1\kappa_2\kappa_3$.

The rate in Theorem 3 is uninformative, as it is specific to the selected implementation. (Some preliminary results and variations can be found in the Supplementary Material, Section 7)

As in Algorithm 3, Algorithm 4 may bypass the approximation of some conditionals. In particular, if $\pi(\alpha \mid \mu)$ can be simulated from and $\pi(\mu \mid x^*, \alpha)$ cannot, we prove in the Supplementary Material that the limiting distribution of our algorithm is the same as the vanilla Approximate Bayesian computation algorithm. On the other hand, if we can simulate from $\pi(\mu \mid x^*, \alpha)$ and not from $\pi(\alpha \mid \mu)$, a version of Theorem 3 is established under more stringent conditions in the Supplementary Material.

3.2. Comparison with vanilla ABC

Recall that in practice the tolerance is provided by an empirical quantile of the distance distribution at each call of an approximate conditional. This means that at each iteration $N_\alpha$ and $N_\mu$, simulations are produced from the conditional prior predictives, respectively, and only the simulation associated with the smallest distance is stored. In Section 7 we explore some further variations in this implementation. The R code used for all simulations can be found at https://github.com/GClarte/ABCG.

We strive to provide a fair comparison between ABC-Gibbs and vanilla ABC and hence aim at using the same overall number of simulations. Since $N_{\text{tot}} = N(N_\alpha + N_\mu)$ is the total number of simulations in Algorithm 4, ABC-Gibbs based on $N$ iterations is naturally comparable with vanilla ABC based on $N_{\text{tot}}$ total simulations.
Fig. 1. ABC estimations of the posterior mean (first column) and variance (second column) of the parameter $\mu_1$ (first row) and the hyperparameter $\alpha$ (second row) for the model given in (1), with $N_{\text{tot}} = 2.1 \cdot 10^4$ and $N_\alpha = N_\mu$ varying from 1 to 400. In all simulations, $N = \lfloor N_{\text{tot}}/N_\mu \rfloor$.

The true value is represented by the horizontal line.

The toy example underlying this comparison is a Normal–Normal model inspired from Gelman et al. (2013),

$$
\mu_j \sim \mathcal{N}(\alpha, \sigma^2), \quad x_{j,k} \sim \mathcal{N}(\mu_j, \sigma^2), \quad j = 1, \ldots, n, \quad k = 1, \ldots, K
$$

with the variance $\sigma^2$ known, and a hyperprior $\alpha \sim \mathcal{U}[-4, 4]$. It is straightforward to check that the assumptions of Theorem 3 hold for this model. Figure 1 illustrates the behaviour of both algorithms, for $\sigma = 1$, $K = 10$, $n = 20$. The true value of the hyperparameter and first parameter are $\alpha = 1.7$ and $\mu_1 = 3.47$, respectively.

The statistic used at both parameter and hyperparameter levels is the corresponding empirical mean and hence it is sufficient. We keep $N_{\text{tot}}$ constant and increase $N_\alpha = N_\mu$ while decreasing $N$. The variability of the estimators deduced from the ABC-Gibbs algorithm reaches a minimum around $N_\alpha = 35$. This is due to a competition between the Monte Carlo error associated by the number of points, which explains the final increase of variance, and the burn-in stage of the Gibbs sampler, which is responsible for the decrease in variance at the beginning.

This toy experiment exhibits a considerable improvement of the parameter estimator when using ABC-Gibbs. This is easily explained by the difficulty for ABC to find a suitable value of $\mu \in \mathbb{R}^{20}$; poor estimation of the parameter ensues. In fact, ABC produces the same output as a non-hierachical model where the $\mu_j$’s are integrated out.
Fig. 2. Comparison of the sampled densities from ABC and ABC-Gibbs, left $\mu_1$, right $\alpha$, the dot-dash line corresponds to the true posterior.

Fig. 3. Evolution of the estimated posterior density of the hyper-parameter $\alpha$ resulting from an increase in $N_\mu$, left $\mu_1$ right $\alpha$. The true posterior is indicated by the dashed line.

The estimated densities of the marginal posteriors produced by both algorithms are provided in Figure 2, which shows that the ABC-Gibbs density is quite far from the true posterior, albeit closer than the ABC, especially for the parameter $\mu_1$. By contrast, the upper level of the hierarchy seems to be poorly explored by ABC. While all graphs and estimates are based on a single simulated dataset, the improvement brought by ABC-Gibbs is found to repeatedly occur over simulations.

To further the comparison, we ran a different experiment, this time with the posterior sample size fixed at $N = 250$. We illustrate the impact of the choice of both $\varepsilon_\alpha$ and $\varepsilon_\mu$ in Figure 3. As $N_\alpha = N_\mu$ increases, that is as the tolerances decrease, the estimated densities concentrate. The density tends to the true posterior, which is to be expected since the approximations are based on sufficient statistics.

4. APPLICATION: MOVING AVERAGE MODEL

4.1. Model and implementation

In this section, we study a hierarchical moving average model. A graphical representation of the hierarchy is shown in Figure 4. We denote $MA_2(\mu, \sigma^2)$ the distribution of a second order moving average model with parameters $\mu = (\mu_1, \mu_2)$ and $\sigma^2$, that is:

$$x(t) = y_t + \mu_1 y_{t-1} + \mu_2 y_{t-2}, \quad \text{with } y_t \sim \mathcal{N}(0, \sigma^2) \text{ for integer } t \geq -1.$$
Dirichlet distributions: using the pseudo-distance $w$ where $q$ standard half-Cauchy distribution, the fact that these quantities appear to have undefined mean and variance.

Third, the $\alpha \rightarrow (\beta_{1,j} - \beta_{2,j}, 2(\beta_{1,j} + \beta_{2,j}) - 1) = (\mu_{j,1}, \mu_{j,2})$, where $(\beta_{1,j}, \beta_{2,j}, \beta_{3,j}) \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3)$, and, if $E$ denotes the exponential distribution and $C_+$ the standard half-Cauchy distribution,

$\alpha = (\alpha_1, \alpha_2, \alpha_3) \sim E(1) \otimes 3, \quad \varsigma = (\varsigma_1, \varsigma_2) \sim C_+ \otimes 2$.

We define $w(x_j)$ the distance between the first two autocorrelations of $x_j$ and $x_j^*$:

$$w^2(x_j) = (\rho_1(x_j) - \rho_1(x_j^*))^2 + (\rho_2(x_j) - \rho_2(x_j^*))^2,$$

and

$$\bar{x}_j = \frac{1}{[\frac{T}{3}]} \sum_{t=1}^{[\frac{T}{3}]} x_j(3t), \quad v(x_j) = \frac{1}{[\frac{T}{3}]} \sum_{t=1}^{[\frac{T}{3}]} \sum_{t=1}^{[\frac{T}{3}]} (x_j(3t) - \bar{x}_j)^2 - \sum_{t=1}^{[\frac{T}{3}]} (x_j^*(3t) - \bar{x}_j^*)^2,$$

where $T$ is the length of the time series. The rationale is that for a $\mathcal{M}_A^2$ model $x(t)$ and $x(t + 3)$ are independent. Vanilla ABC uses a related single pseudo-distance defined by

$$\delta(x) = \sum_{j=1}^{n} \left\{ \frac{w(x_j)}{q_j} + \frac{v(x_j)}{q_j'} \right\}, \quad (3)$$

where $q_j$ and $q_j'$ are the 0.1% quantiles of $w(x_j)$ and $v(x_j)$, respectively. This choice is constrained by the fact that these quantities appear to have undefined mean and variance.

For the current model, we implemented the full version of ABC-Gibbs. First, the $\mu_{j}$’s are updated using the pseudo-distance $w(x_j)$. Second, the update of $\alpha$ relies on the sufficient statistic associated with Dirichlet distributions:

$$\mu \rightarrow \left( \sum_j \log((\mu_{j,2} + 2\mu_{j,1} + 1)/4), \sum_j \log((\mu_{j,2} + 2\mu_{j,1} + 1)/4) - \mu_{j,1} \right)$$

Third, the $\sigma_j$’s are conditioned on the pseudo-distance $v(x_j)$. And last, $\varsigma$ is updated using the standard sufficient statistic associated with gamma distributions.

The two algorithms output samples from the two pseudo-posteriors. To compare the quality of the two samplers, we simulate new synthetic data from each parameter set in the output, and compute the distance (3) between observed and simulated samples, which is the distance used by ABC. If ABC-Gibbs produces a smaller value than the ABC associated with this distance, this is an indicator of a better fit of the ABC-Gibbs distribution with the true posterior. As in the previous experiment, the total number of simulations of the time series is used as the default measure of the computational cost for the associated algorithm.
Fig. 5. For the toy dataset of subsection 4.2, approximate posterior of $\mu_1$ compared with the prior for ABC-Gibbs (left) and ABC (right).

4.2. Toy dataset

Consider a synthetic dataset of $n = 5$ times series each with length $T = 100$. Both samplers return samples of size $N = 1000$. The hyperparameters used to produce the true parameters and the simulated observed series are $\alpha = (1, 2, 3)$ and $\zeta = (1, 1)$. In ABC-Gibbs, the $\mu_j$’s are updated based on $N_\mu = 1000$ time series, while the other parameters are updated based on $N_\alpha = N_\zeta = N_\sigma = 100$ replicas. The overall computational cost for ABC-Gibbs is $N_{\text{tot}} = 5.5 \times 10^6$, also used by ABC to run $1.1 \times 10^6$ simulations of the whole hierarchy. The computational cost is slightly superior for ABC, as we have to simulate many more Dirichlet and Gamma random variables.

When evaluating the expectation of the posterior predictive distance (3), ABC-Gibbs achieves an average of $274.1 \pm 2.5$, and ABC reaches an average of $436.8 \pm 1.6$, based on 100 replicates. The sample produced by ABC-Gibbs thus offers a noticeably better quality than the one produced by ABC from this perspective. The ABC output barely differs from a simulation from the prior, as shown in Figure 5 for the parameter $\mu$.

4.3. Stellar flux

We now apply this model to stellar flux data. The 8GHz daily flux emitted by seven stellar objects is analysed in Lazio et al. (2008), and the data were made public by the Naval Research Laboratory: https://tinyurl.com/yxorvl4u. Once a few missing observations have been removed, Lazio et al. (2008) suggest that the model described in Section 4.1 may be well suited to these data, with $T = 208$. In ABC-Gibbs, the $\mu_j$’s are updated based on $N_\mu = 500$ time series, while the other parameters require $N_\alpha = N_\zeta = N_\sigma = 100$ replicas. (The overall computing time is the same for the toy and the current datasets; one hour on an Intel Xeon CPU E5-2630 v4 with rate 2.20GHz.)

The average posterior distance to the observed sample is $232.8 \pm 1.25$ for ABC-Gibbs and $535 \pm 0.95$ for ABC. The poor fit of the latter is confirmed in Figure 6, as it again stays quite close to the prior for the $\mu$’s. Since our model differs from the one proposed in Lazio et al. (2008), estimators cannot be directly compared.

5. Approximate Bayesian Gibbs Sampling with Full Dependence

The concept of ABC-Gibbs is by no means restricted to hierarchical settings. It applies in full generality to any decomposition or completion of the parameter $\theta$ into $n$ terms, $(\theta_1, \ldots, \theta_n)$. For simplicity’s sake, we only analyse below the case of $n = 2$ parameters, and furthermore assume that $\theta_1$ and $\theta_2$ are a priori independent. The extension to $n \geq 2$ parameters, or non-independent parameters, is straightforward though cumbersome. The generic Algorithm 3 and Theorem 1 can thus be adapted to non-hierarchical models where $\theta = (\theta_1, \theta_2)$, such that the conditional posteriors $\pi(\theta_1 \mid x^*, \theta_2)$ and $\pi(\theta_2 \mid x^*, \theta_1)$ depend on the entire dataset $x^*$ rather than a significantly smaller subset. This setting implies that the approx-
imulation steps in ABC-Gibbs will mostly require the simulation of objects of the same size as in ABC. However, efficiency may be gained from using component-wise steps and different summary statistics \( s_1 \) and \( s_2 \) for \( \theta_1 \) and \( \theta_2 \), resulting in an alternative version of ABC-Gibbs, Algorithm 5.

**Algorithm 5.** Full dependence ABC-Gibbs sampler

**Input:** observed dataset \( x^\star \), number of iterations \( N \), starting point \( \theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}) \), threshold \( \varepsilon = (\varepsilon_1, \varepsilon_2) \), summary statistics \( s_1 \) and \( s_2 \)

**Output:** sample \( (\theta^{(1)}, \ldots, \theta^{(N)}) \)

for \( i = 1, \ldots, N \) do

\[
\begin{align*}
\theta_1^{(i)} &\sim \pi_{\varepsilon_1}(\cdot \mid x^\star, s_1, \theta_1^{(i-1)}) \\
\theta_2^{(i)} &\sim \pi_{\varepsilon_2}(\cdot \mid x^\star, s_2, \theta_2^{(i-1)})
\end{align*}
\]

When the statistics \( s_1 \) and \( s_2 \) are identical, a single distance can be used, with \( \varepsilon_1 = \varepsilon_2 \). The resulting stationary distribution is then the same as for ABC, since it is proportional to

\[
\int \pi(\theta_1)\pi(\theta_2)f(x \mid \theta_1, \theta_2)1_{y(s_1(x), s_2(x^\star)) \leq \varepsilon_1} \, dx.
\]

Formally, these statistics should, however, be different, since more efficient and smaller-dimension statistics can be calibrated to each parameter. The following alternative to Theorem 1 applies:

**Theorem 4.** Assume that

\[
\begin{align*}
\kappa_1 &= \inf_{\theta_1, \theta_2} \pi(B_{s_1(x^\star), \varepsilon_1} \mid \theta_1, \theta_2) > 0, \\
\kappa_2 &= \inf_{\theta_1, \theta_2} \pi(B_{s_2(x^\star), \varepsilon_2} \mid \theta_1, \theta_2) > 0, \\
\kappa_3 &= \sup_{\theta_1, \hat{\theta}_1, \theta_2} \| \pi(\cdot \mid \theta_1, \theta_2) - \pi(\cdot \mid \hat{\theta}_1, \theta_2) \|_{TV} < 1/2.
\end{align*}
\]

Then the Markov chain produced by Algorithm 5 geometrically converges in total variation distance to a stationary distribution \( \nu_\varepsilon \) with geometric rate \( 1 - \kappa_1 \kappa_2 (1 - 2\kappa_3) \).

As an illustration, consider an example inspired by inverse problems (Kaipio & Fox, 2011), in a simplified version. Let \( y \) be the solution of the heat equation on a circle defined for \( (\tau, z) \in [0, T] \times [0, 1] \) by

\[
\partial_\tau y(z, \tau) = \partial_z (\theta(z) \partial_z y(z, \tau)) ,
\]
with \( \theta(z) = \sum_{j=1}^{n} \theta_j \mathbf{1}_{[\frac{j-1}{n}, \frac{j}{n})}(z) \) and with boundary conditions
\[
y(z,0) = y_0(z), \quad y(0, \tau) = y(1, \tau)
\]

We assume \( y_0 \) known and the parameter is \( \theta = (\theta_1, \ldots, \theta_n) \). The equation is discretized towards its numerical resolution. For this purpose, the first order finite elements method relies on discretization steps of size \( \frac{1}{n} \) for \( z \) and \( \Delta \) for \( \tau \). A stepwise approximation of the solution is thus \( \hat{y}(z, t) = \sum_{j=1}^{n} y_{j,t} \phi_j(z) \), where, for \( j < n \),
\[
\phi_j(z) = (1 - |nz - j|) \frac{1}{|nz - j| < 1}
\]
and \( \phi_n(z) = (1 - nz) \frac{1}{0 < z < 1/n} + \frac{(nz - n + 1)}{1 - 1/n < z < 1} \), and with \( y_{j,t} \) defined by
\[
\frac{y_{j,t+1} - y_{j,t}}{3\Delta} + \frac{y_{j+1,t+1} - y_{j+1,t}}{6\Delta} + \frac{y_{j-1,t+1} - y_{j-1,t}}{6\Delta} = y_{j,t+1} (\theta_{j+1} + \theta_j) - y_{j-1,t+1} \theta_j - y_{j+1,t+1} \theta_{j+1}.
\]

We then observe a noisy version of this process, chosen as \( x_{j,t} = N(\hat{y}_{j,t}, \sigma^2) \).

In ABC-Gibbs, each parameter \( \theta_m \) is updated with summary statistics the observations at locations \( m - 2, m - 1, m, m + 1 \). ABC relies on the whole data as statistic. In the experiments, \( n = 20 \) and \( \Delta = 0.1 \), with a prior \( \theta_j \sim U[0, 1] \), independently.

We compared the two methods, using as above the same simulation budget. We performed several experiments for both algorithms with various values of \( N_{\epsilon} \), keeping the total number of simulations constant at \( N_{\text{tot}} = N_{\epsilon} \cdot N = 8 \cdot 10^6 \). As \( N_{\epsilon} \) increases, the size \( N \) of the posterior sample decreases. Figure 7 illustrates the estimations of \( \theta_1 \). The ABC-Gibbs estimate is much closer to the true value of the parameter \( \theta_1 = 0.75 \), with a smaller variance.

As in the previous sections, ABC-Gibbs is much more efficient than ABC. For instance, Figure 8 shows that the posterior sample of \( \theta_1 \) is more peaked around the true value for ABC-Gibbs. We repeated this experiment for a wide range of values for \( \theta \). In all our experiments, ABC-Gibbs gave estimates close to the true value, and was never outperformed by ABC. This is confirmed by evaluating the expectation of the posterior predictive distance to the whole data, ABC-Gibbs achieves an average of \( 39.2 \pm 0.002 \), and ABC reaches an average of \( 103.8 \pm 0.002 \), based on 100 replicates.

6. Compatibility of the approximate marginals

6.1. Necessary and sufficient compatibility condition for model (1)

In Section 3, addressing model (1), we gave conditions in Theorem 3 for Algorithm 4 to have a limiting distribution \( \nu_{\epsilon} \). However, the nature of this distribution is unknown. We also showed that as the toler-

Fig. 7. For the heat equation model, mean and variance of the ABC and ABC-Gibbs estimators of \( \theta_1 \) as \( N_{\epsilon} \) increases, selected from among 20 parameters. The horizontal line shows the true value of \( \theta_1 \).
ance parameter $\varepsilon$ decreases, $\nu_\varepsilon$ tends to the stationary distribution $\nu_0$ of a Gibbs sampler with generators $\pi(\alpha)\pi(s_\alpha(\mu) | \alpha)$ and $\pi(\mu | \alpha)\pi(s_\mu(x) | \mu)$. It is possible that these generators are incompatible, that is, that there is no joint distribution associated with them. In such settings, the stationary distribution $\nu_0$ does not enjoy these generators as conditionals. The incompatibility of conditionals may seem contradictory with the fact that our algorithm does converge to a distribution, but in the case of a compact parameter space there always exist a limiting distribution, the main issue being rather that the limiting distribution has no straightforward Bayesian interpretation.

In fact, the approximate conditionals cannot be compatible when $s_\alpha$, the statistic for the update of the hyperparameter $\alpha$, is not sufficient. According to Arnold & Press (1989) a necessary and sufficient condition for the conditionals to be compatible is the existence of two measurable functions $u(\alpha)$ and $v(\mu)$, such that
\[
\frac{\pi(\alpha)\pi(s_\alpha(\mu) | \alpha)}{\pi(\mu | \alpha)f(s_\mu(x) | \mu)} = u(\alpha)v(\mu),
\]
that is, there exist $\tilde{u}$ and $\tilde{v}$ such that $\pi(\mu | \alpha) = \pi(s_\alpha(\mu) | \alpha)\tilde{u}(\alpha)\tilde{v}(\mu)$. This occurs if and only if $s_\alpha$ is sufficient, by the factorization theorem.

On the positive side, there is no compatibility issue when $s_\alpha$ is sufficient, whatever the statistic $s_\mu$ may be. In particular, there is no compatibility issue in the special case of an exact simulation from $\pi(\alpha | \mu)$. In such settings, the limiting distribution is then $\pi(\alpha)\pi(\mu | \alpha)f(s_\mu(x) | \mu)$, which is the same as with ABC when the tolerance is set to zero and the summary statistic is $s_\mu$.

6.2. Switching target distribution and algorithm

Our method thus produces ABC algorithms that are not necessarily built on a genuine joint distribution, although they may still converge to a stationary distribution. This feature is detrimental to the validation of the approach in that establishing convergence may prove delicate, and making inferential sense of the limiting distribution may prove impossible. In this section, we examine a different version of our method that is an exact Gibbs sampler for a bona fide joint distribution inspired by ABC-Gibbs. Furthermore, this joint distribution can be precisely characterised when the tolerance goes to zero.

Consider the case of the hierarchical model (1) (with $n = 1$). The successive generations of $\alpha^{(i)}$ and $\mu^{(i)}$ imply the generation of auxiliary variables $\tilde{\alpha}^{(i)}$ and $\tilde{x}$, respectively produced from the conditionals $\pi(\mu | \alpha^{(i)})$ and $\pi(x | \mu^{(i)})$, until the tolerance conditions are met. There exists an alternative version that enjoys a joint distribution, defined as follows
\[
\tilde{\pi}(\alpha, \mu | x^*) \propto \pi(\alpha)q(\mu) \int \pi(\tilde{\mu} | \alpha)\mathbf{1}_{\eta(s_\alpha(\mu), s_\mu(\tilde{\mu})) < \varepsilon_\alpha} d\tilde{\mu} \int f(\tilde{x} | \mu)\pi(x^* | \mu)\mathbf{1}_{\eta(s_\mu(x^*), s_\mu(\tilde{x})) < \varepsilon_\mu} d\tilde{x},
\]
with \( q \) an arbitrary distribution on \( \mu \). It induces full conditionals given by

\[
\tilde{\pi}(\alpha \mid \mu) \propto \pi(\alpha) \int \pi(\mu' \mid \alpha) \mathbf{1}_{\eta(s_\alpha(\mu), s_\alpha(\mu')) < \varepsilon_\alpha} \, d\tilde{x}
\]

(4)

and

\[
\tilde{\pi}(\mu \mid \alpha, x^*) \propto \pi(\mu) \int \pi(\mu' \mid \alpha) \mathbf{1}_{\eta(s_\alpha(\mu), s_\alpha(\mu')) < \varepsilon_\alpha} \int f(x^* \mid \mu) \pi(x^* \mid \mu) \mathbf{1}_{\eta(s_\mu(x^*), s_\mu(\tilde{x})) < \varepsilon_\mu} \, d\tilde{x}.
\]

(5)

The first conditional is simulated by prior simulations of \( \alpha \sim \pi(\alpha) \) and of \( \tilde{\mu} \sim \pi(\tilde{\mu} \mid \alpha) \) until \( \eta(s_\alpha(\mu), s_\alpha(\tilde{\mu})) < \varepsilon_\alpha \). The second conditional can be simulated by a prior simulation of \( \mu \) by first simulating from \( \pi(\alpha) \) and second simulating from \( \pi(\mu \mid \alpha) \), and of both auxiliary variables \( \tilde{\mu} \) and \( \tilde{x} \) until both conditions are satisfied. This alternative is coherent in the sense that the conditionals are now compatible with the joint. Furthermore

\textbf{Proposition 1.} \textit{When both} \( \varepsilon_\alpha \) \textit{and} \( \varepsilon_\mu \) \textit{decrease to zero, the Gibbs sampler associated with the conditionals (4) and (5) is stationary for the posterior distribution proportional to}

\[
\pi(\alpha, \mu) q(s_\alpha(\mu)) f(s_\mu(x^*) \mid \mu)
\]

that is, for the likelihood associated with \( s_\mu(x^*) \) and the prior distribution proportional to \( \pi(\alpha, \mu) q(s_\alpha(\mu)) \).

This algorithm is slightly more computationally costly. However, as \( q \) is arbitrary, it can be chosen as a closed-form density and hence we can weight the final sample by an importance weight of \( 1/q(s_\alpha(\mu)) \) to get back an interpretable posterior. In particular, if \( q \) is chosen as a Uniform density, which may require a reparameterisation onto a compact set, the outcome is distributed from this posterior. This implies this version of ABC-Gibbs can be corrected towards a simulation from the posterior with the same approximation error as the vanilla ABC algorithm.

7. Discussion

The curse of dimensionality remains the major jamming block for the expansion of ABC methodology to more complex models as most of its avatars see their cost rise with the dimensions of the parameter and of the data (Li & Fearnhead, 2018). This is particularly the case for high-dimensional parameters, since they require summary statistics that are at least of the same dimension and, unless the model under study is amenable to easily computed estimates of these parameters, a much larger collection of statistics is usually unavoidable. Breaking this curse of dimensionality by Gibbs-like steps is thus as important for ABC methods as it was for Monte Carlo methods (Gelfand & Smith, 1990) as relying on a small number of summary statistics facilitates the derivation of automated or semi-automated approaches (Fearnhead & Prangle, 2012; Raynal et al., 2019) and offers the potential for simulating pseudo-data of much smaller size. In appropriate settings, ABC-Gibbs sampling provides a noticeable improvement of the efficiency of approximate Bayesian computation methods. We have established some sufficient conditions for the convergence of ABC-Gibbs algorithms. Questions remain, from checking such conditions in practice to a better understanding of the limiting distributions from an inferential viewpoint. The proposal made in Section 6 is a step in this direction, as it allows for a closed-form description of the limiting distribution. Alternative proposals that are more appropriate for the particular dependence structure of the model under study should be devised. Constructing or selecting a low-dimension informative summary statistic for the approximation of the conditionals might be an unavoidable challenge to further improve the quality of the results.
ACKNOWLEDGEMENTS

This paper greatly benefited from early discussions with Anthony Ebert, Kerrie Mengersen, and Pierre Pudlo, to whom we are most grateful. We are also grateful to the Jean Morlet Chair which partly supported meeting in the Centre International de Rencontres Mathématiques in Luminy.

REFERENCES

ARNOLD, B. C. & PRESS, S. J. (1989). Compatible conditional distributions. Journal of the American Statistical Association 84, 152–156.

BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M. & ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. Biometrika 96(3), 759–774.

BEAUMONT, M. A., ZHANG, W. & BALDING, D. J. (2002). Approximate Bayesian Computation in Population Genetics. Genetics 162, 2025–2035.

CARLIN, B. & LOUIS, T. (1996). Bayes and Empirical Bayes Methods for Data Analysis. London: Chapman and Hall.

CUILÉREY, K., BLUM, M. G. B., GAGGIOTTI, O. E. & FRANCOIS, O. (2010). Approximate Bayesian computation (ABC) in practice. Trends in ecology & evolution 25, 410–418.

FEARNHEAD, P. & PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 74, 419–474.

FRAZIER, D., MARTIN, G., ROBERT, C. & ROUSSEAU, J. (2018). Asymptotic properties of approximate Bayesian computation. Biometrika 105, 593–607.

GELFAND, A. E. & SMITH, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association 85, 398–409.

GELMAN, A., CARLIN, J. B., SHERFLAND, A., DUNSON, D. B., VEHTARI, A. & RUBIN, D. B. (2013). Bayesian Data Analysis, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

GELMAN, S. & GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 6, 721–741.

KAIPIO, J. P. & FOX, C. (2011). The Bayesian Framework for Inverse Problems in Heat Transfer. Heat Transfer Engineering 32, 718–753.

LAZIO, T. J. W., BALDWIN, E., D. GHIGO, F., FIEDLER, R. S., FOSTER, R. K. & JOHNSTON, A. (2008). A Dual-Frequency, Multiyear Monitoring Program of Compact Radio Sources. The Astrophysical Journal Supplement Series 166, 265.

LI, W. & FEARNHEAD, P. (2018). On the asymptotic efficiency of approximate Bayesian computation estimators. Biometrika 105, 285–299.

LINDLEY, D. & SMITH, A. (1972). Bayes estimates for the linear model. Journal of the Royal Statistical Society 34, 1–41.

LUNN, D., THOMAS, A., BEST, N. & SPiegelhalter, D. (2010). The BUGS Book: A Practical Introduction to Bayesian Analysis. New York: Chapman & Hall/CRC Press.

MARIN, J.-M., PUDLO, P., ROBERT, C. P. & RYDER, R. J. (2012). Approximate Bayesian computational methods. Statistics and Computing 22, 1167–1180.

MARJORAM, P., MOLITOR, J., PLAGNOL, V. & TAVARE, S. (2003). Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences 100, 15324–15328.

MEYR, S. P. & TWEDEDE, R. L. (1993). Markov Chains and Stochastic Stability. London: Springer-Verlag.

MOORES, M. T., DROVANDI, C. C., MENGERSSEN, K. & ROBERT, C. P. (2015). Pre-processing for approximate Bayesian computation in image analysis. Statistics and Computing 25, 23–33.

NUMMELIN, E. (1978). A splitting technique for Harris recurrent chains. Zeit. Warsch. Verv. Gebiete 43, 309–318.

RAYNAL, L., MARIN, J.-M., PUDLO, P., RIBATET, M., ROBERT, C. P. & ESTOPPE, A. (2019). ABC random forests for Bayesian parameter inference. Bioinformatics 35, 1720–1728.

ROBERT, C. & CASELLA, G. (2004). Monte Carlo Statistical Methods. New York: Springer Verlag, 2nd ed.

SISSON, S., YAN, Y. & BEAUMONT, M., eds. (2018). Handbook of Approximate Bayesian Computation. New York: Chapman and Hall/CRC.

TAVARE, S., BALDING, D. J., GRIFFITHS, R. C. & DONNELLY, P. (1997). Inferring Coalescence Times From DNA Sequence Data. Genetics 145, 505–518.

TONI, T., WELCH, C., STREWKOWA, N., IPSEN, A. & STUMPF, M. P. H. (2008). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface 6, 187–202.
Supplementary material

In this supplementary material, we define $\Theta_j$ as the space where $\theta_j$ lives. For the proofs that pertain to model (1) we define $\mathcal{A}$ as the space in which $\alpha$ lives and $\mathcal{B}$ as the one where $\mu$ lives. For a space $E$, $\mathcal{P}(E)$ is the space of the probability distributions over $E$.

7.1. Generalities on total variation distance

The main tool in our proofs is the total variation distance used by Nummelin (1978); Meyn & Tweedie (1993), which is defined for $\nu$ and $\tilde{\nu}$, two probability distributions over the same space, $E$, as

$$||\nu - \tilde{\nu}||_{TV} = \frac{1}{2} \inf_{\gamma \in \Gamma(\nu, \tilde{\nu})} \text{Pr}\{x \neq y \mid (x, y) \sim \gamma\},$$

where $\Gamma(\nu, \tilde{\nu})$ is the set of all couplings between $\nu$ and $\tilde{\nu}$ that is of all the probability distributions $\gamma$ on $E \times E$ such that $\int \gamma(x, y)dx = \nu$ and $\int \gamma(x, y)dy = \tilde{\nu}$. To handle this distance, we build explicit coupling between the distributions to find an upper bound on this distance. Notice that there always exists an optimal coupling between two distributions, that is a coupling $\gamma_0$ such that $||\nu - \tilde{\nu}||_{TV} = \frac{1}{2}\text{Pr}\{x \neq y \mid (x, y) \sim \gamma_0\}$

7.2. Proof of Theorem 1

In this proof, we drop the conditionings on $x^*$, $s_1$, and $s_2$, as they have no use in the computations and create a notational burden.

We only have to prove that the Markov chain $(\theta_i^{(t)})_{1 \leq i \leq N}$ has a stationary distribution. We show that $Q : \mathcal{P}(\Theta_1) \rightarrow \mathcal{P}(\Theta_1)$ the mapping associated with the transition kernel is a contraction, that is, we prove that $||Q\nu - Q\tilde{\nu}||_{TV} \leq L||\nu - \tilde{\nu}||_{TV}$, for some $0 < L < 1$ for all measures $\nu$ and $\tilde{\nu}$.

To build a coupling between $Q\nu$ and $Q\tilde{\nu}$ we define a coupling kernel $Q : \mathcal{P}(\Theta_1 \times \Theta_1) \rightarrow \mathcal{P}(\Theta_1 \times \Theta_1)$, taking a coupling $\xi_0$ as argument, such that $\int Q\xi_0(x, y)dx = Q\nu$ and $\int Q\xi_0(x, y)dy = Q\tilde{\nu}$. This coupling kernel is explicitly defined by the following procedure, taking $(\theta_1, \tilde{\theta}_1) \sim \xi_0$, a coupling of $\nu$ and $\tilde{\nu}$ as input and returning $(\theta_1', \tilde{\theta}_1') \sim Q\xi_0$:

Algorithm 6. Coupling procedure

**Input:** $(\theta_1, \tilde{\theta}_1) \sim \xi_0$.

**Output:** $(\theta_1', \tilde{\theta}_1') \sim Q\xi_0$.

$(\theta_2, \tilde{\theta}_2) \sim \xi_1(\cdot \mid \theta_1, \tilde{\theta}_1)$;

$(\theta_1', \tilde{\theta}_1') \sim \xi_2(\cdot \mid \theta_2, \tilde{\theta}_2)$.

For $\xi_1(\cdot \mid \theta_1, \tilde{\theta}_1)$ be an optimal coupling between $\pi_{\varepsilon_2}(\cdot \mid \theta_1)$ and $\pi_{\varepsilon_2}(\cdot \mid \tilde{\theta}_1)$, and $\xi_2(\cdot \mid \theta_2, \tilde{\theta}_2)$ be an optimal coupling between $\pi_{\varepsilon_1}(\cdot \mid \theta_2)$ and $\pi_{\varepsilon_1}(\cdot \mid \tilde{\theta}_2)$. This procedure verifies that if $\theta_1 = \tilde{\theta}_1$ then $\theta_1' = \tilde{\theta}_1'$, as for any distribution, $\nu_0$ and itself is $(x, y) \mapsto \nu_0(x)\delta_{x=y}$.

In the proofs we need to choose $\xi_0$ as the optimal coupling between $\nu$ and $\tilde{\nu}$. In the following, $\tilde{\gamma} = Q\xi_0$

$$||Q\nu - Q\tilde{\nu}||_{TV} = \frac{1}{2} \inf_{\gamma \in \Gamma(Q\nu, Q\tilde{\nu})} \text{Pr}\{\theta_1' \neq \tilde{\theta}_1' \mid (\theta_1, \tilde{\theta}_1) \sim \gamma\} \leq \frac{1}{2}\text{Pr}\{\theta_1' \neq \tilde{\theta}_1' \mid \theta_1 \neq \tilde{\theta}_1, (\theta_1', \tilde{\theta}_1') \sim \tilde{\gamma}\} \text{Pr}_{\xi_0}(\theta_1 = \tilde{\theta}_1) + \frac{1}{2}\text{Pr}\{\theta_1' \neq \tilde{\theta}_1' \mid \theta_1 \neq \tilde{\theta}_1, (\theta_1', \tilde{\theta}_1') \sim \tilde{\gamma}\} \text{Pr}_{\xi_0}(\theta_1, \tilde{\theta}_1) \leq \frac{1}{2}\text{Pr}\{\theta_1' \neq \tilde{\theta}_1' \mid \theta_1 \neq \tilde{\theta}_1, (\theta_1', \tilde{\theta}_1') \sim \tilde{\gamma}\} \text{Pr}_{\xi_0}(\theta_1, \tilde{\theta}_1) \leq ||\nu - \tilde{\nu}||_{TV} \text{Pr}_{\xi_0}\{\theta_1' \neq \tilde{\theta}_1' \mid \theta_1, \tilde{\theta}_1\sim \tilde{\gamma}\} = 1 - \text{Pr}_{\xi}\{\theta_1' = \tilde{\theta}_1' \mid \theta_1 \neq \tilde{\theta}_1, (\theta_1', \tilde{\theta}_1') \sim \tilde{\gamma}\},$$

that is to find a lower bound on the probability that two different value $\theta_1$ and $\tilde{\theta}_1$ transition to the same value.
If \( \theta_2 = \hat{\theta}_2 \) then necessarily, \( \theta'_1 = \hat{\theta}'_1 \), in other words, if the coupling is successful at the first step of the procedure it is sufficient. Meaning that we can bound from below the coupling probability by the coupling probability for the first step of the procedure \( \theta_2 \).

\[
\text{pr}\{ \theta'_1 = \hat{\theta}'_1 \mid \theta_1 \neq \hat{\theta}_1, (\theta'_1, \hat{\theta}'_1) \sim \hat{\gamma} \} \geq 1 - 2\| \pi_\epsilon (\cdot \mid \theta_1) - \pi_\epsilon (\cdot \mid \hat{\theta}_2) \|_{TV} > 1 - 2\kappa > 0.
\]

This proves that the map \( \nu \mapsto Q\nu \) is a contraction. The space of all measures on \( \mathcal{A} \) is complete when endowed with the total variation distance. Furthermore the subspace of all probability distributions on \( \Theta_1 \) is stable by \( Q \). Hence, according to the Banach fixed point theorem it enjoys a fixed point and in particular the sequence \( (Q^n\pi) \), with \( \pi \) an arbitrary prior distribution, converges to this fixed point with rate \( 1 - 2\kappa \).

7.3. Proof of Theorem 2

In this proof, we need a coupling between two chains with different transition kernels. Let \( \nu_\epsilon \) be the target distribution of the approximate Gibbs sampler and \( \nu_0 \) be the target distribution of the exact Gibbs sampler. Let \( (\theta_1, \hat{\theta}_1) \) be a realisation of an optimal coupling \( \xi_0 \) between \( \nu_\epsilon \) and \( \nu_0 \). As before we propose a coupling procedure:

**Algorithm 7. Coupling procedure**

**Input:** \( (\theta_1, \hat{\theta}_1) \sim \xi_0 \).

**Output:** \( (\theta'_1, \hat{\theta}'_1) \sim Q\xi_0 \).

\( (\theta_2, \hat{\theta}_2) \sim \xi_3 (\cdot \mid \theta_1, \hat{\theta}_1) \);

\( (\theta'_1, \hat{\theta}'_1) \sim \xi_4 (\cdot \mid \theta_2, \hat{\theta}_2) \).

with \( \xi_3 (\cdot \mid \theta_1, \hat{\theta}_1) \) an optimal coupling between \( \pi_\epsilon (\cdot \mid \theta_1) \) and \( \pi (\cdot \mid \hat{\theta}_1) \), and \( \xi_4 (\cdot \mid \theta_2, \hat{\theta}_2) \) an optimal coupling between \( \pi_\eta (\cdot \mid \theta_2) \) and \( \pi (\cdot \mid \hat{\theta}_2) \).

As the distributions \( \nu_\epsilon \) and \( \nu_0 \) are stationary for the evolution process, we have

\[
\text{pr}(\theta'_1 \neq \hat{\theta}'_1) = \text{pr}(\theta'_1 \neq \hat{\theta}'_1 \mid \theta_1 \neq \hat{\theta}_1) \text{pr}(\theta_1 \neq \hat{\theta}_1) + \text{pr}(\theta'_1 \neq \hat{\theta}'_1 \mid \theta_1 = \hat{\theta}_1) \text{pr}(\theta'_1 = \hat{\theta}'_1) \leq \frac{\text{pr}(\theta'_1 \neq \hat{\theta}'_1 \mid \theta_1 = \hat{\theta}_1)}{\text{pr}(\theta'_1 = \hat{\theta}'_1 \mid \theta_1 = \hat{\theta}_1)}.
\]

As before we roughly bound the denominator:

\[
\text{pr}(\theta'_1 = \hat{\theta}'_1 \mid \theta_1 \neq \hat{\theta}_1) \geq (1 - 2\kappa) \sup_{\theta_1, \hat{\theta}_1} \| \pi_\epsilon (\cdot \mid \theta_1) - \pi (\cdot \mid \hat{\theta}_1) \|_{TV} \geq 1 - 2L_0.
\]

For the numerator we have, with \( \theta_2 \) and \( \hat{\theta}_2 \) the transitory values of the second parameter,

\[
\text{pr}(\theta'_1 \neq \hat{\theta}'_1 \mid \theta_1 = \hat{\theta}_1) \leq \text{pr}(\theta'_1 \neq \hat{\theta}'_1 \mid \theta_2 = \hat{\theta}_2) \text{pr}(\theta_2 \neq \hat{\theta}_2 \mid \theta_1 = \hat{\theta}_1) + \text{pr}(\theta'_1 \neq \hat{\theta}'_1 \mid \theta_2 = \hat{\theta}_2) \text{pr}(\theta_2 \neq \hat{\theta}_2 \mid \theta_1 = \hat{\theta}_1) \leq \sup_{\theta_2} \text{pr}(\theta'_1 \neq \hat{\theta}_1 \mid (\theta_1, \hat{\theta}_1) \sim \xi_4 (\cdot \mid \theta_2, \hat{\theta}_2)) + \sup_{\theta_2} \text{pr}(\theta_2 \neq \hat{\theta}_2 \mid (\theta_2, \hat{\theta}_2) \sim \xi_3 (\cdot \mid \theta_1, \hat{\theta}_1)) \leq 2L_1(\epsilon_1) + 2L_2(\epsilon_2)
\]

Putting together both estimates gives the bound of the theorem.

7.4. Generalisation of Theorem 1

We can extend the convergence result to the general case \( n > 2 \):
Theorem 5. Assume that for all \( \ell \)

\[
\kappa_\ell = \sup_{\theta > \ell} \sup_{\theta < \ell} \| \pi_{\varepsilon, \ell}(. \mid x^*, s_\ell, \theta < \ell, \theta \geq \ell) - \pi_{\varepsilon, \ell}(. \mid x^*, s_\ell, \theta < \ell, \tilde{\theta} > \ell) \|_{TV}
\]

with \( \theta > \ell = (\theta_{\ell+1}, \theta_{\ell+2}, \ldots, \theta_n) \), and \( \theta < \ell = (\theta_1, \theta_2, \ldots, \theta_{\ell-1}) \). Then, the Markov chain produced by Algorithm 3 geometrically converges in total variation distance to a stationary distribution \( \nu_\varepsilon \) with geometric rate \( 1 - \prod \kappa_\ell \).

The proof of this theorem is a straightforward adaptation of the previous proof, with the same coupling procedure. The condition comes from the fact that in this procedure we sequentially try to couple each \( \theta_\ell \) using the previous one, already coupled.

7.5. Proofs that are specific to the hierarchical case

In the particular case of Model 1 we provide specific convergence results. They are based on a particular implementation of ABC-Gibbs, presented for \( n = 1 \) and in the case of an analytically available conditional density \( \pi(\mu \mid \alpha, x^*) \), in Algorithm 8. We will gradually weaken the hypothesis to finally prove Theorem 3.

Algorithm 8. Implementation of ABC-Gibbs used in the proofs.

\[\text{Input:} \alpha^{(0)} \sim \pi(\alpha), \mu^{(0)} \sim \pi(\mu \mid \alpha^{(0)}) \]

\[\text{Output:} \text{A sample } (\alpha^{(i)}, \mu^{(i)})_{1 \leq i \leq N} \]

\[\text{for } i = 1, \ldots, N \text{ do} \]

\[\mu^c \sim \pi(. \mid \alpha^{(i-1)}, x^*) \]

\[\alpha^c \sim \pi \]

\[\tilde{\mu} \sim \pi(. \mid \alpha^c) \]

\[\text{if } \eta(s_\alpha(\tilde{\mu}), s_\alpha(\mu^c)) < \varepsilon_\alpha \text{ then} \]

\[\mu^{(i+1)} \leftarrow \mu^c \]

\[\alpha^{(i+1)} \leftarrow \alpha^c \]

\[\text{else} \]

\[\mu^{(i+1)} \leftarrow \mu^{(i)} \]

\[\alpha^{(i+1)} \leftarrow \alpha^{(i)} \]

First we state the most restrictive result:

Theorem 6. Assume that the following conditions are all satisfied

\[
\kappa_1 = \inf_{\mu} \pi(B_{s_\alpha(\mu), \varepsilon_\alpha / 4} > 0
\]

\[
\kappa_2 = \inf_{\alpha} \inf_{\mu} \pi(B_{s_\alpha(\mu), 3\varepsilon_\alpha / 2} \mid \alpha, x) > 0.
\]

Then, the Markov chain associated with Algorithm 8 enjoys an invariant distribution, and it converges geometrically to this invariant measure with rate \( 1 - \kappa_1 \kappa_2 \) for the total variation distance.

Proof. The technique of the proof is essentially similar to the previous one. Let \( \nu \) and \( \tilde{\nu} \) be two distributions over \( A \). We describe the evolution of \( \alpha, \tilde{\alpha} \) into \( \alpha', \tilde{\alpha}' \), though the kernel \( \tilde{Q} \). We denote \( \mu, \tilde{\mu} \) the transitory second parameter.
Algorithm 9. Coupling procedure

**Input:** (α, ā)

**Output:** (α', ā')

if α ≠ ā then
   | (μ, ě) ~ π(· | α, x) ⊗ π(· | ā, x)
else
   | μ = ě ~ π(μ | α, x)
   | α' ~ π(· | α)
   | μ' ~ π(· | α')
if η(s_α(μ), s_α(μ')) ≤ ε_α then
   | α' ← α'
else
   | ā' ← ā

This process defines a transition kernel $\hat{Q}$ for two coupled chains. As in the previous proofs, if α = ā then α' = ā'.

Let (α, ā) ~ $ξ$, an optimal coupling between ν and $\tilde{ν}$. Then,

$$
\|Qν - Q\tilde{ν}\|_{TV} = \frac{1}{2} \inf_{γ ∈ P(ν, Q\tilde{ν})} \text{pr}\{α' ≠ ā' | (α', ā') ~ γ\}
\leq \frac{1}{2} \text{pr}_ξ\{α' ≠ ā' | α = ā, (α', ā') ~ \hat{Q}ξ\} \text{pr}_ξ(α = ā)
+ \frac{1}{2} \text{pr}_ξ\{α' ≠ ā' | α ≠ ā, (α', ā') ~ \hat{Q}ξ\} \text{pr}_ξ(α ≠ ā)
\leq \frac{1}{2} \text{pr}_ξ\{α' ≠ ā' | α ≠ ā, (α', ā') ~ \hat{Q}ξ\} \text{pr}_ξ(α ≠ ā)
\leq \|ν - \tilde{ν}\|_{TV} \text{pr}_ξ(α' ≠ ā' | α ≠ ā, (α', ā') ~ \hat{Q}ξ).
$$

It is sufficient to find a uniform upper bound on $\text{pr}_ν,\tilde{ν}(α' ≠ ā' | α ≠ ā, (α', ā') ~ \hat{Q}ξ) = \int \pi(s_α(μ) ≠ ā', α ≠ ā, (α', ā') ~ \hat{Q}ξ) = \int \pi(\nu(α') ≠ ā' | α ≠ ā, (α', ā') ~ \hat{Q}ξ) = \int \nu(α)\tilde{ν}(\tilde{α})d\tilde{α}dμd\tilde{μ}$. Notice that we can choose our coupling $ξ$ such that conditionally on α ≠ ā the marginals are independent.

$$
\text{pr}_ν,\tilde{ν}(α' ≠ ā' | α ≠ ā) = \int \pi\{B_{s_α(μ), ε_α} \cap B_{s_α(μ'), ε_α}\} \pi(μ | α, x)
× \pi(μ' | ā, x)\nu(α)\tilde{ν}(\tilde{α})d\tilde{α}dμd\tilde{μ}
= \int \pi\{B_{s_α(μ), ε_α} \cap B_{s_α(μ'), ε_α}\} \pi(μ | α, x)
× \pi(μ' | ā, x)\nu(α)\tilde{ν}(\tilde{α})1_{[η(s_α(μ), s_α(μ')) ≤ 3ε_α/2]}d\tilde{α}dμd\tilde{μ}
+ \int \pi\{B_{s_α(μ), ε_α} \cap B_{s_α(μ'), ε_α}\} \pi(μ | α, x)
× \pi(μ' | ā, x)\nu(α)\tilde{ν}(\tilde{α})1_{[η(s_α(μ), s_α(μ')) > 3ε_α/2]}d\tilde{α}dμd\tilde{μ}
= I_1 + I_2.
$$
We now bound $I_1$ and $I_2$.

\[
I_1 = \int \pi \left\{ (B_{s,\alpha}(\mu), \varepsilon) \cap B_{s,\alpha}(\bar{\mu}), \varepsilon \right\} \pi(\mu | \alpha, x)\pi(\bar{\mu} | \bar{\alpha}, x)\nu(\alpha, \bar{\alpha})1_{\{\eta(\alpha, \bar{\alpha}) \leq 3\varepsilon/2\}} \text{d}\alpha \text{d}\alpha \text{d}\mu \text{d}\bar{\mu}
\]

\[
\leq \int \pi(\eta(\alpha, \bar{\alpha}) \leq 3\varepsilon/2) \pi(\mu | \alpha, x)\pi(\bar{\mu} | \bar{\alpha}, x)\nu(\alpha, \bar{\alpha})1_{\{\eta(\alpha, \bar{\alpha}) \leq 3\varepsilon/2\}} \text{d}\alpha \text{d}\alpha \text{d}\mu \text{d}\bar{\mu}
\]

\[
\leq \text{pr}_{\nu, \bar{\nu}} \{ \eta(\alpha, \bar{\alpha}) \leq 3\varepsilon/2 \}
\]

Finally, putting both inequalities together, we have $I_1 + I_2 \leq 1 - \kappa_1\kappa_2$, with $\kappa_1\kappa_2 > 0$, and

\[
\|Q\nu - Q\bar{\nu}\|_{TV} \leq (1 - \kappa_1\kappa_2)\|\nu - \bar{\nu}\|_{TV}.
\]

The conclusion is the same as in the previous proof.

**Remark 1.** In the proof, when we describe the coupling kernel, we generate $\mu$ and $\bar{\mu}$ independently if $\alpha$ and $\bar{\alpha}$ are different, and as a single $\mu$ if they are equal. This is a particular coupling of the distributions $\pi(\cdot | \alpha, x)$ and $\pi(\cdot | \bar{\alpha}, x)$ Here, the link between Theorem 1 and this one becomes clear, as we make the coupling explicit toward reaching a bound in total variation.

The first relaxation of hypothesis we can do is removing the compacity of $A$: the resulting theorem is theorem 3.

**Proof.** With the same notations as before, we merely need to find a lower bound:

\[
I_3 = \int \text{pr}(\mu \in B_{s,\alpha}(\mu), \varepsilon/2) \pi(\mu | \alpha, x)\pi(\bar{\mu} | \bar{\alpha}, x)\alpha(\alpha, \bar{\alpha})1_{\{\eta(\alpha, \bar{\alpha}) > 3\varepsilon/2\}} \text{d}\alpha \text{d}\alpha \text{d}\mu \text{d}\bar{\mu}
\]

\[
I_3 \geq \int \text{pr}(B_{s,\alpha}(\mu), \varepsilon/2) \pi(\mu | \alpha, x)\pi(\bar{\mu} | \bar{\alpha}, x)\nu(\alpha, \bar{\alpha})1_{\{\eta(\alpha, \bar{\alpha}) > 3\varepsilon/2\}} \text{d}\alpha \text{d}\alpha \text{d}\mu \text{d}\bar{\mu}
\]

\[
\geq \text{pr}(\mu \in B_{s,\alpha}(\mu), \varepsilon/2) \pi(\mu | \alpha, x)\pi(\bar{\mu} | \bar{\alpha}, x)\nu(\alpha, \bar{\alpha})1_{\{\eta(\alpha, \bar{\alpha}) > 3\varepsilon/2\}} \text{d}\alpha \text{d}\alpha \text{d}\mu \text{d}\bar{\mu}
\]

\[
\geq \kappa_1\kappa_2\varepsilon^2
\]

as the convexity of $C$ ensures that $1_{\{\eta(\alpha, \bar{\alpha}) > 2\varepsilon/2\}} \geq 1_{\eta(\alpha, \bar{\alpha}) \in C}1_{\eta(\alpha, \bar{\alpha}) \in C}$. \qed

To remove the compacity hypothesis on $B$, we need a stronger hypothesis:
THEOREM 7. Assume that there exist $\mathcal{H} \subset \mathcal{P}(A)$ stable by $Q$ and $A \subset \mathcal{A}$ and $C \subset s_\alpha(B)$ with finite measure such that:

\[
\kappa_1 = \inf_{s_\alpha(\mu) \in C} \pi(B_{s_\alpha(\mu), \varepsilon_\alpha/4}) > 0;
\]

\[
\kappa_2 = \inf_{\alpha \in A} \inf_{s_\alpha(\mu) \in C} \pi(B_{s_\alpha(\mu), 3\varepsilon_\alpha/2} | \alpha, x) > 0;
\]

\[
\kappa_3 = \inf_{\alpha \in A} \pi(s_\alpha(\mu) \in C | \alpha, x) > 0;
\]

\[
\kappa_4 = \inf_{\nu \in \mathcal{H}} \nu(A) > 0.
\]

Then, the Markov chain associated with Algorithm 8 enjoys an invariant distribution, and it converges geometrically to this invariant measure with rate $1 - \kappa_1 \kappa_2 \kappa_2^2 \kappa_4^2$.

Proof. Similarly to previous proofs, we have

\[
I_3 \geq \int \text{pr}(\mu \in B_{s_\alpha(\mu), \varepsilon_\alpha/2}) \pi(\mu | \alpha, x) \pi(\tilde{\alpha}, x) \nu(\alpha) \nu(\tilde{\alpha})
\]
\[
\times \mathbf{1}_{s_\alpha(\mu) \in C} \mathbf{1}_{s_\alpha(\tilde{\mu}) \in C} \mathbf{1}_{[ \eta(s_\alpha(\mu), s_\alpha(\tilde{\mu})), 3\varepsilon_\alpha/2]} \, d\alpha d\tilde{\alpha} d\mu d\tilde{\mu}
\]
\[
\geq \int \kappa_1 \kappa_2 \kappa_2^2 \kappa_4^2 \nu(\alpha) \nu(\tilde{\alpha}) \mathbf{1}_{\alpha \in A} \mathbf{1}_{\tilde{\alpha} \in A}
\]
\[
\geq \kappa_1 \kappa_2 \kappa_2^2 \kappa_4^2.
\]

7.6. Full dependence approximate Gibbs sampling

The proof of Theorem 4 is as follows.

We propose the following coupling, with $\xi(\cdot | \theta_1, \tilde{\theta}_1, \theta_2) \tilde{\theta}_2)$ is an optimal coupling between $f(\cdot | \theta_1, \theta_2)$ and $f(\cdot | \theta_1, \tilde{\theta}_2)$:

Algorithm 10. Coupling procedure

Input: $(\theta_1, \tilde{\theta}_1)$

Output: $(\theta_1', \tilde{\theta}_1')$

$\theta_2' \sim \pi(\theta_2)$;

$(x, \tilde{x}) \sim \xi(x, \tilde{x} | \theta_1, \tilde{\theta}_1, \theta_2', \tilde{\theta}_2')$

$\theta_1' \sim \pi(\theta_1)$

$x^c \sim \pi(x^c | \theta_1', \theta_2')$

if $\eta(s_2(x), s_2(x^*)) < \varepsilon_2$, $\eta(s_1(x^c), s_1(x^c)) < \varepsilon_1$ then

$\theta_1' \leftarrow \theta_1$

else

$\tilde{\theta}_1' \leftarrow \tilde{\theta}_1$

if $\eta(s_2(\tilde{x}), s_2(x^*)) < \varepsilon_2$, $\eta(s_1(x^c), s_1(x^c)) < \varepsilon_1$ then

$\tilde{\theta}_1' \leftarrow \tilde{\theta}_1$

else

$\tilde{\theta}_1' \leftarrow \tilde{\theta}_1$
As before, if $\theta_1 = \tilde{\theta}_1$ then $\theta' = \tilde{\theta}'$, thus we want an upper bound for:

$$\text{pr}(\theta'_1 \neq \tilde{\theta}'_1 \mid \theta_1 \neq \tilde{\theta}_1)$$

$$= 1 - \int \pi(\theta^c_2)\pi(\theta^c_1)\pi(x^c \mid \theta^c_1, \theta^c_2)1_{\eta(s_2(x), s_2(x^*)) < \epsilon_2}1_{\eta(s_2(\tilde{x}), s_2(x^*)) < \epsilon_2}1_{\eta(s_1(x^c), s_1(x^*)) < \epsilon_1} \nu(\theta_1) \bar{\nu}(\tilde{\theta}_1)$$

$$\times d\xi(x, \tilde{x} \mid \theta_1, \tilde{\theta}_1, \theta^c_2)\text{d}\theta^c_1\text{d}\tilde{\theta}_1\text{d}\theta^c_2$$

$$\leq 1 - \int \pi(\theta^c_2)\pi(\theta^c_1)\min_{\tilde{\theta}=\theta_1, \tilde{\theta}_1} \pi(B_{s_2(x^*)}, \epsilon_2 \mid \theta, \theta^c_2)\pi(B_{s_1(x^*)}, \epsilon_1 \mid \theta^c_1, \theta^c_2)$$

$$\times d\xi(x, \tilde{x} \mid \theta_1, \tilde{\theta}_1, \theta^c_2)\nu(\tilde{\theta}_1)\nu(\theta_1)\bar{\nu}(\tilde{\theta}_1)\text{d}\theta^c_1\text{d}\tilde{\theta}_1\text{d}\theta^c_2$$

$$\leq 1 - \kappa_1\kappa_2 \int \pi(\theta^c_2)\pi(\theta^c_1)\{1 - 2\|\pi(\cdot \mid \theta_1, \theta^c_2) - \pi(\cdot \mid \tilde{\theta}_1, \theta^c_2)\|_{TV}\} \nu(\theta_1)\nu(\tilde{\theta}_1)\text{d}\theta^c_1\text{d}\tilde{\theta}_1\text{d}\theta^c_2$$

$$\leq 1 - \kappa_1\kappa_2(1 - \kappa_3).$$

where we used the optimality of the coupling $\xi$. 