Statistics and Computing (2019) 29:449–482
https://doi.org/10.1007/s11222-018-9817-3

Informed sub-sampling MCMC: approximate Bayesian inference for large datasets

Florian Maire[1,2] · Nial Friel[1,2] · Pierre Alquier[3]

Received: 26 June 2017 / Accepted: 4 June 2018 / Published online: 9 June 2018

Abstract
This paper introduces a framework for speeding up Bayesian inference conducted in presence of large datasets. We design a Markov chain whose transition kernel uses an unknown fraction of fixed size of the available data that is randomly refreshed throughout the algorithm. Inspired by the Approximate Bayesian Computation literature, the subsampling process is guided by the fidelity to the observed data, as measured by summary statistics. The resulting algorithm, Informed Sub-Sampling MCMC, is a generic and flexible approach which, contrary to existing scalable methodologies, preserves the simplicity of the Metropolis–Hastings algorithm. Even though exactness is lost, i.e the chain distribution approximates the posterior, we study and quantify theoretically this bias and show on a diverse set of examples that it yields excellent performances when the computational budget is limited. If available and cheap to compute, we show that setting the summary statistics as the maximum likelihood estimator is supported by theoretical arguments.

Keywords Bayesian inference · Big-data · Approximate Bayesian Computation · noisy Markov chain Monte Carlo

Mathematics Subject Classification Primary 65C40, 65C60; Secondary 62F15

1 Introduction

The development of statistical methodology that scale to large datasets represents a significant research frontier in modern statistics. This paper presents a generic and flexible approach to directly address this challenge when a Bayesian strategy is followed. Given a set of observed data $(Y_1, \dots, Y_N)$, a specified prior distribution p and a likelihood function f, estimating parameters $\theta \in \Theta$ of the model proceeds via exploration of the posterior distribution π defined on $(\Theta, \mathcal{B}(\Theta))$ by

$$\pi(\mathrm{d}\theta \mid Y_1, \dots, Y_N) \propto f(Y_1, \dots, Y_N \mid \theta) p(\mathrm{d}\theta) . \qquad (1)$$

Stochastic computation methods such as Monte Carlo methods allow one to estimate characteristics of π. In Bayesian inference, Markov chain Monte Carlo (MCMC) methods remain the most widely used strategy. Paradoxically, improvements in data acquisition technologies together with increased storage capacities, present a new challenge for these methods. Indeed, the size of the data set N (along with the dimension of each observation) can become so large, that even a routine likelihood evaluation is made prohibitively computationally intensive. As a consequence, MCMC methods such as the Metropolis–Hastings algorithm (Metropolis et al. 1953) cannot be considered for reasonable runtime. This issue has recently generated a lot of research activity, see Bardenet et al. (2017) for a comprehensive review.

Most of the scalable MCMC methods proposed in the literature are based on approximations of the Metropolis–Hastings (M–H) algorithm. In the sequel, we will refer to as *exact approximations*, algorithms that produce samples from the target distribution when the chain is in the stationary regime, as opposed to *approximate* methods that do not. Central to those scalable MCMC approaches is the idea that only the calculation of the likelihood of a subset of data would be required to simulate a new state of the Markov chain.

✉ Florian Maire
florian.maire@ucd.ie

[1] School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

[2] The Insight Centre of Data Analytics, University College Dublin, Dublin, Ireland

[3] CREST, ENSAE, Université Paris Saclay, Paris, France

Following the development of pseudo-marginal algorithms (Andrieu and Roberts 2009; Andrieu and Vihola 2015), a first direction has been to replace the likelihoods in the M–H acceptance ratio by positive unbiased estimators (based on a subset of data). Although appealing since exact, this approach remains (for now) mostly theoretical because such estimators are in general not available (Jacob et al. 2015). Attempts to circumvent the positivity and unbiasedness requirements of the estimator have been studied in Quiroz et al. (2016) and Quiroz et al. (2015), respectively. In both cases, the authors resort to sophisticated control variates, which can be computationally expensive to compute.

Other authors have proposed to approximate the log-likelihood ratio by subsampling data points (Korattikara et al. (2014); Bardenet et al. (2014, 2017)), the objective being to mimic the accept/reject decision that would be achieved by the Metropolis–Hastings algorithm. Even though the resulting algorithms are not exact, the *Confidence sampler* proposed in Bardenet et al. (2014) and refined in Bardenet et al. (2017) is designed such that the accept/reject decision is, with an arbitrarily high probability, identical to that taken by the Metropolis–Hastings algorithm. The construction of this algorithm, based on concentration inequalities, allows to bound the L1 distance between the stationary distribution of the algorithm and π. The price to pay is that the number of likelihood evaluations is not fixed but adaptively set by the algorithm at each iteration and, as noted in Bardenet et al. (2017), it is of order $\mathcal{O}(N)$ when the chain reaches equilibrium. This number can be brought down if an accurate proxy of the log-likelihood ratio, acting as control variates, is available, as demonstrated in Bardenet et al. (2017).

More recently, a stream of research has shed light on the use of continuous time Markov processes (Zig-Zag process, Langevin diffusion) to perform Bayesian analysis of tall dataset (Bierkens et al. 2018; Pollock et al. 2016; Fearnhead et al. 2016). The computational bottleneck for this class of methods is the calculation of the gradient of the log-likelihood and it has been shown that provided that an unbiased estimate of this gradient is used, they remain exact. Here again, the use of control variates to reduce the variance of the estimator is in practice essential to reach the full potential of these methods. However, we note that those approaches represent a significant departure from the M–H algorithm and as such lose its implementational simplicity.

In this paper, we propose Informed Sub-Sampling MCMC (ISS-MCMC), a novel methodology which aims to make the best use of a computational resource available for a given computational run-time, while still preserving the celebrated simplicity of the standard M–H sampler. The state space Θ is extended with an n-dimensional vector of unique integers $U_k \subset \{1,\ldots,N\}$ identifying a subset of the data used by the Markov transition kernel at the the k-th iteration of the algorithm, where $n \ll N$ is set according to the available computational budget. Central to our approach is the fact that each subset is weighted according to a *similarity measure* with respect to the full set of data through summary statistics, in the spirit of Approximate Bayesian Computation (ABC) (see e.g. Marin et al. 2012). The subset variable is randomly refreshed at each iteration according to the similarity measure. The Markov chain transition kernel only uses a fraction n/N of the available data which is by construction –and contrary to Maclaurin and Adams (2015), Korattikara et al. (2014) and Bardenet et al. (2014)– held constant throughout the algorithm. Moreover, unlike most of the papers mentioned before, our method can be applied to virtually any model (involving *i.i.d.* data or not), as it does not require any assumption on the likelihood function nor on the prior distribution. Our algorithm can be cast as a *noisy* MCMC method since the marginal in θ of our Markov chain targets an approximation of π that we quantify using the framework established in Alquier et al. (2016). In the special case where the data are *i.i.d.* realizations from an exponential model, we prove that when the summary statistics is set as the sufficient statistics, this yields an optimal approximation, in the sense of minimizing an upper bound of the Kullback-Leibler (KL) divergence between π and the marginal target of our method. In the general case, we show that setting the summary statistics as the maximum likelihood estimator allows to bound the approximation error (in L1 distance) of our algorithm. We connect our work to a number of recent papers including Rudolf and Schweizer (2018); Huggins and Zou (2016); Dalalyan (2017) that bound approximation error of MCMC algorithms, using the Wasserstein metric.

To summarize, the main contribution of our work is to show that, under verifiable conditions, it is possible to infer π through a scalable approximation of the M–H algorithm where the computational budget of each iteration is fixed (through the subset size n). To do so, it is necessary to draw the subsets according to a similarity measure with respect to the full data set and not uniformly at random, as previously explored in the literature. We show that setting the similarity measure as the squared L2 distance between the full dataset and subsample maximum likelihood estimators is supported by theoretical arguments.

Section 2 presents a striking real data example which we hope will help the reader to understand the problem we address and motivate the solution we propose, without going into further technical details at this stage. In Sect. 3, we provide theoretical results concerning exponential-family models, which we illustrate through a probit example. This section allows us to justify our motivations supporting the Informed Sub-Sampling general methodology which is rigorously presented in Sect. 4. In Sect. 5, we study the transition kernel of our algorithm and show that it yields a Markov chain targeting, marginally, an approximation of π. The approximation error is quantified and we provide theoretical justifications for setting up the Informed Sub-Sampling tuning parameters, including the choice of summary statistics. Finally, in Sect. 6, our method is used to estimate parameters of an autoregressive time series and a logistic regression model. It is also illustrated to perform a binary classification task. We compare the performance of our algorithm with the Stochastic Gradient Langevin Dynamic (SGLD) from Welling and Teh (2011) and two versions of Subsampled Likelihoods Bardenet et al. (2014, 2017).

2 An introductory example

We showcase the principles of our approach on a first real data example. The problem at hand is to infer some template shapes of handwritten digits from the MNIST database (http://yann.lecun.com/exdb/mnist/).

Example 1 The data $Y_1, Y_2, \ldots$ are modelled by a deformable template model (Allassonnière et al. 2007). Each data Y_i is a 15×15 pixel image representing an handwritten digit whose conditional distribution given its class $J(i) \in (0, 1, \ldots, 9)$ is a random deformation of the template shape, parameterized by a $d = 256$ dimensional vector $\theta_{J(i)}$. Assuming small deformations, the model is similar to a standard regression problem:

$$Y_i = \phi(\theta_{J(i)}) + \sigma^2 \epsilon_i\,, \tag{2}$$

where Y_i is regarded as a vector $\mathbb{R}^{225}$, $\phi : \mathbb{R}^{256} \to \mathbb{R}^{225}$ is some deterministic mapping and $\sigma > 0$ is the standard deviation of the additive noise $\epsilon_i \sim \mathcal{N}(0_{225}, \mathrm{Id}_{225})$.

Given a set of N labeled images $Y_1, Y_2, \ldots, Y_N$ and a prior distribution for $\theta = \{\theta_1, \ldots, \theta_9\}$, one can estimate θ through its posterior distribution π, for example using the Metropolis–Hastings (M–H) algorithm (Metropolis et al. 1953). However, since the regression function ϕ in (2) is quite sophisticated, even a single likelihood evaluation is expensive to calculate. As a result, the M–H efficiency can be questioned as computing the N likelihoods in the M–H ratio dramatically slows down each transition.

At this stage, we do not provide precise details on the Informed Sub-Sampling MCMC method but we simply provide an insight of the rationale of our approach. It designs a Markov chain whose transition kernel targets a scaled version of the posterior distribution of the parameter of interest θ given a random subset of n images ($n \ll N$). More specifically, we inject in the standard M–H transition a decision about *refreshing* the subset of data, which, as a result, will change randomly over time. In this example, we use the knowledge of the observation labels to promote subsets of images in which the proportion of each digit is balanced.

We consider $N = 10{,}000$ images of five digits $1, \ldots, 5$, subsets of size $n = 100$ and a non-informative Gaussian prior for θ, as specified in Allassonnière et al. (2007). Figure 1 indicates a striking advantage of our method compared to a standard M–H using the same $N = 10{,}000$ images. In this scenario, we allow a fixed computational budget (1 h) for both methods and compare the estimation of the mean estimate of the two Markov chains. Qualitatively, the upper part of Fig. 1 compares the estimated template shapes of the five digits at different time steps and shows that ISS-MCMC allows one to extract template shapes much quicker than the standard M–H, while still reaching an apparent similar graphical quality after 1 h. This fact is confirmed quantitatively, in the lower part of Fig. 1, which plots, against time and for both methods, the Euclidean distance between the Markov chain mean estimate and the maximum likelihood estimate $(\theta_1^*, \ldots, \theta_5^*)$ obtained using a stochastic EM (Allassonnière et al. 2007). More precisely, we compare the real valued function $\{d(t),\ t \in \mathbb{R}\}$ defined as

$$d(t) = \sum_{j=1}^{5} \|\theta_j^* - \mu(\theta_{j,1:\kappa(t)})\|, \qquad \text{where}$$

$$\begin{cases} \forall t \in \mathbb{R},\ \kappa(t) = \max_{k \in \mathbb{N}}\{t \geq \tau_k\}\,, \\ \tau_k \text{ is the time at the end of the } k \text{-th iteration}\,, \\ \forall k \in \mathbb{N},\ \mu(\theta_{j,1:k}) = (1/k)\sum_{\ell=1}^{k} \theta_{j,\ell}\,, \end{cases}$$

where we have defined for $(j, k) \in \{1, \ldots, 5\} \times \mathbb{N}, \theta_{j,k}$ as the j-th class parameter obtained after k iterations of the Markov chains. For a vector $x \in \mathbb{R}^n$, $\|\cdot\|$ will refer to the usual L2 norm on $\mathbb{R}^n$, unless stated otherwise.

One can see that the transient phase of the Informed Sub-Sampling Markov chain is significantly shorter than that of the Metropolis–Hastings chain. More details on Example 1 can be found at Sect. 6. In particular, Fig. 14 shows that the stationary distribution of Informed Sub-Sampling matches reasonably well π, which is a primary concern in Bayesian inference.

Our algorithm provides very encouraging results for this real data example. We motivate and formalize our method in Sects. 3 and 4 and provide theoretical arguments supporting it at Sect. 5.

3 Approximation of the posterior distribution in exponential models: an optimality result

In this section, we consider the case of N independent and identically distributed (*i.i.d.*) observations from an exponential model. Sampling from the posterior distribution of such models using the Metropolis–Hastings algorithm is effortless since the information conveyed by the N observations

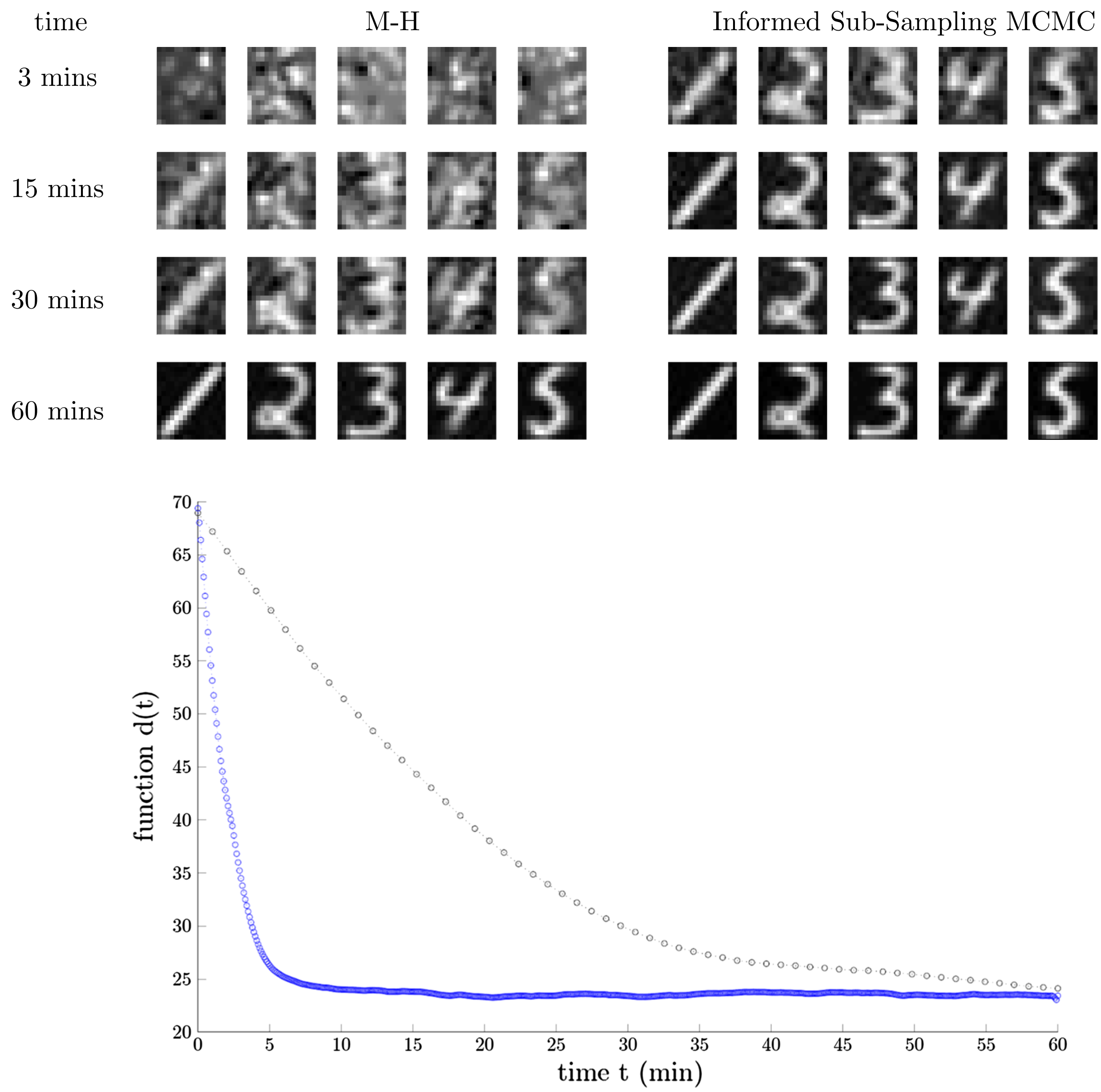


Fig. 1 (Example 1: Handwritten digits) Efficiency of template estimation through M–H (black) and Informed Sub-Sampling MCMC (blue). (Color figure online)

is contained in the sufficient statistics vector, which needs to be calculated only once.

The existence of sufficient statistics in this type of models allows us to establish a number of theoretical results that will be used to design and justify our Informed Sub-Sampling methodology that approximately samples from posterior distributions in general contexts, i.e non-*i.i.d.* observations from general likelihood models without sufficient statistics. More precisely, Propositions 1 and 2 put forward an optimal approximation of the posterior distribution π by a distribution $\tilde{\pi}_n$ of the parameter of interest given only a subsample of n observations. Finally, Proposition 3 justifies the introduction of a probability distribution on the set of subsamples. This is an essential element of our work as it represents a significant departure from all existing subsampling methodologies proposed in the Markov chain Monte Carlo literature, that have assumed uniform distribution on the subsamples.

3.1 Notation

Let $(Y_1, \ldots, Y_N) \in \mathsf{Y}^N$ be a set of *i.i.d.* observed data ($\mathsf{Y} \subseteq \mathbb{R}^m$, $m > 0$) and define

- $Y_{i:j} = (Y_i, \ldots, Y_j)$ if $1 \leq i \leq j \leq N$ with the convention that $Y_{i:j} = \{\emptyset\}$, otherwise.
- $Y_U = \{Y_k,\ k \in U\}$, where $U \subseteq \{1, \ldots, N\}$.

In this section, we assume that the likelihood model f belongs to the exponential family and is fully specified by a vector of parameters $\theta \in \Theta$, ($\Theta \subseteq \mathbb{R}^d$, $d > 0$), a bounded mapping $g : \Theta \to \mathsf{S}$ and a sufficient statistic mapping $S : \mathsf{Y} \to \mathsf{S}$ ($\mathsf{S} \subseteq \mathbb{R}^s, s > 0$) such that

$$f(y \mid \theta) = \exp\left\{g(\theta)^T S(y)\right\} \Big/ L(\theta)\,,$$
$$L(\theta) = \int_{Y \in \mathsf{Y}} \exp\left\{S(y)^T g(\theta)\right\} \mathrm{d}y\,,$$

is the density of the likelihood distribution with respect to the Lebesgue measure. The posterior distribution π is defined on the measurable space $(\Theta, \mathcal{B}(\Theta))$ by its density function

$$\pi(\theta \mid Y_{1:N}) = p(\theta) \frac{\exp\left\{\sum_{k=1}^N S(Y_k)^T g(\theta)\right\}}{L(\theta)^N} \Big/ Z(Y_{1:N})\,, \tag{3}$$

where

$$Z(Y_{1:N}) = \int p(\mathrm{d}\theta) \frac{\exp\left\{\sum_{k=1}^N S(Y_k)^T g(\theta)\right\}}{L(\theta)^N}\,. \tag{4}$$

In Eq. (3), p is a prior distribution defined on $(\Theta, \mathcal{B}(\Theta))$ and with some abuse of notation, p denotes also the probability density function on Θ.

For all $n \leq N$, we define U_n as the set of possible combinations of n different integer numbers less than or equal to N and $\mathcal{U}_n$ as the powerset of U_n. In the sequel, we set n as a constant and wish to compare the posterior distribution π (3) with any distribution from the family $\mathsf{F}_n = \{\tilde{\pi}_n(U),\ U \in \mathsf{U}_n\}$, where for all $U \in \mathsf{U}_n$, we have defined $\tilde{\pi}_n(U)$ as the distribution on $(\Theta, \mathcal{B}(\Theta))$ with probability density function

$$\tilde{\pi}_n(\theta \mid Y_U) \propto p(\theta) f(Y_U \mid \theta)^{N/n}\,. \tag{5}$$

3.2 Optimal subsets for the Kullback–Leibler divergence between π and $\tilde{\pi}_n$

Recall that for two measures π and $\tilde{\pi}$ defined on the same measurable space $(\Theta, \mathcal{B}(\Theta))$, the Kullback-Leibler (KL) divergence between π and $\tilde{\pi}$ is defined as:

$$\mathrm{KL}(\pi, \tilde{\pi}) = \mathbb{E}_\pi \left\{\log \frac{\pi(\theta)}{\tilde{\pi}(\theta)}\right\}\,. \tag{6}$$

Although not a proper distance between probability measures defined on the same space, $\mathrm{KL}(\pi, \tilde{\pi})$ is used as a similarity criterion between π and $\tilde{\pi}$. It can be interpreted in information theory as a measure of the information lost when $\tilde{\pi}$ is used to approximate π, which is our primary concern here. We now state the main result of this section.

Proposition 1 *For any subset $U \in \mathsf{U}_n$, define the vector of difference of sufficient statistics between the whole dataset and the subset Y_U as*

$$\Delta_n(U) = \sum_{k=1}^N S(Y_k) - (N/n) \sum_{k \in U} S(Y_k)\,. \tag{7}$$

Then, the following inequality holds:

$$KL\{\pi, \tilde{\pi}_n(U)\} \leq B(Y, U)\,, \tag{8}$$

where

$$B(Y, U) = \log \mathbb{E}_\pi \exp\{\|\mathbb{E}_\pi(g(\theta)) - g(\theta)\| \, \|\Delta_n(U)\|\} \tag{9}$$

and $\|\cdot\|$ is the L2 norm.

The proof is detailed in Appendix A.1 and follows from straightforward algebra and applying Cauchy-Schwartz inequality. Note that by definition of B, we remark that for any two subsets $(U_1, U_2) \in \mathsf{U}_n^2$,

$$\|\Delta_n(U_1)\| \leq \|\Delta_n(U_2)\| \;\Rightarrow\; B(Y, U_1) \leq B(Y, U_2)\,.$$

The following corollary is an immediate consequence of Proposition 1.

Corollary 1 *Define the set:*

$$\mathsf{U}_n^\star := \left\{U \in \mathsf{U}_n,\quad \frac{1}{N}\sum_{k=1}^N S(Y_k) = \frac{1}{n}\sum_{k \in U} S(Y_k)\right\}\,. \tag{10}$$

If $\mathsf{U}_n^\star$ is non-empty, then for any $U \in \mathsf{U}_n^\star$, then $\pi(\theta \mid Y) = \tilde{\pi}_n(\theta \mid Y_U)$, π-almost everywhere.

A stronger result can be obtained under the assumption that a Bernstein-von Mises Theorem Van der Vaart (2000); Le Cam (1986) holds for the concentration of π to its Normal approximation:

$$\hat{\pi}(\,\cdot \mid Y_{1:N}) := \mathcal{N}\left(\theta^*(Y_{1:N}), I^{-1}(\theta_0)/N\right)\,, \tag{11}$$

where $\mathcal{N}$ denotes the Normal distribution, $\theta^*(Y_{1:N}) = \arg\max_{\theta \in \Theta} f(Y_{1:N} \mid \theta)$, $\theta_0 \in \Theta$ is some parameter and $I(\theta)$ is the Fisher information matrix given $Y_{1:N}$ at θ.

Proposition 2 *Let $(U_1, U_2) \in \mathsf{U}_n^2$. Assume that for all $i \in \{1, \ldots, d\}$, $|\Delta_n(U_1)^{(i)}| \leq |\Delta_n(U_2)^{(i)}|$, where $|\Delta_n(U_1)^{(i)}|$ refers to the i-th element of $\Delta_n(U_1)$ (7). Then $\widehat{KL}_n(U_1) \leq \widehat{KL}_n(U_2)$, where $\widehat{KL}_n(U)$ is the Kullback-Leibler divergence between the asymptotic approximation of the posterior $\hat{\pi}$ (11) and $\tilde{\pi}_n(U)$ (5).*

The proof is detailed in Appendix A.2. Note that the asymptotic approximation is for $N \to \infty$ and for a fixed n and is thus relevant to the context of our analysis.

3.3 Weighting the subsamples

Consider the distribution $\nu_{n,\epsilon}$ on the discrete space U_n defined for all $\epsilon \geq 0$ by:

$$\nu_{n,\epsilon}(U) \propto \exp\left\{-\epsilon\|\Delta_n(U)\|^2\right\}\,, \qquad \text{for all } U \in \mathsf{U}_n\,. \quad (12)$$

The distribution $\nu_{n,\epsilon}$ assigns a weight to any subset according to their representativeness with respect to the full dataset. When $\epsilon = 0$, $\nu_{n,\epsilon}$ is uniform on U_n while when $\epsilon \to \infty$, $\nu_{n,\epsilon}$ is uniform on the set of subset(s) that minimize(s) $U \mapsto \|\Delta_n(U)\|$. Proposition 2 suggests that for exponential models, the optimal inference based on subsamples of size n is obtained by picking the subposterior $\pi_n(U)$ (5) using the distribution $U \sim \nu_{n,\epsilon}$ with $\epsilon \to \infty$.

We now state Proposition 3. This result is important even though somewhat obscure at this stage. Indeed, we will show that it is a necessary condition for the method we introduce in Sect. 4 to converge. In fact, moving away to general models (i.e non *i.i.d.* and non exponential) amounts to relax the sufficient statistics existence assumption as well as the $\epsilon \to \infty$ condition. This will be achieved by constructing a class of summary statistics for the model at hand for which a similar result to Proposition 3 holds.

Proposition 3 *For any $\theta \in \Theta$ and $\epsilon > 0$, there exists $M < \infty$ such that:*

$$\mathbb{E}_{n,\epsilon}\left\{\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\right\} < M\,, \quad (13)$$

where $\mathbb{E}_{n,\epsilon}$ is the expectation under $\nu_{n,\epsilon}$, as defined in (12).

The proof is postponed to Appendix A.3. Note that Proposition 3 essentially holds because $\log \nu_{n,\epsilon}$ is quadratic in $\|\Delta_n(U)\|$. Other weighting schemes for the subsets (e.g. uniform weights or weights $\propto \exp\{-\epsilon\|\Delta_n(U)\|\}$) would not necessarily allow to bound $\mathbb{E}_{n,\epsilon}\{f(Y\,|\,\theta)/f(Y_U\,|\,\theta)^{N/n}\}$.

3.4 Illustration with a probit model: effect of choice of sub-sample

We consider a pedagogical example, based on a probit model, to illustrate the results from the previous subsections.

Example 2 A probit model is used in regression problems in which a binary variable $Y_k \in \{0, 1\}$ is observed through the following sequence of independent random experiments, defined for all $k \in \{1, \ldots, N\}$ as:

(i) Draw $X_k \sim \mathcal{N}(\theta^*, \gamma^2)$
(ii) Set Y_k as follows

$$Y_k = \begin{cases} 1, & \text{if } X_k > 0,\\ 0, & \text{otherwise.}\end{cases} \quad (14)$$

Table 1 (Example 2: Probit model) Comparison of the KL divergence between π and the optimal $\tilde{\pi}_n \in \mathsf{F}_n$ ($\|\Delta_n(U)\| = 3$) and other distributions in F_n

n	$\|\Delta_n(U)\|$	KL $\{\pi, \tilde{\pi}_n(U)\}$	$B(Y, U)$
1000	3	0.004	0.04
1000	14	0.11	0.18
1000	23	0.19	0.29
100	33	0.41	0.54

Observing a large number of realizations $Y_1, \ldots, Y_N$, we aim to estimate the posterior distribution of θ. If γ is unknown, the model is not identifiable and for simplicity we considered it as known here. The likelihood function can be expressed as

$$\begin{aligned} f(Y_k\,|\,\theta) &= \alpha(\theta)^{Y_k}(1-\alpha(\theta))^{(1-Y_k)}\\ &= (1-\alpha(\theta))\left(\frac{\alpha(\theta)}{1-\alpha(\theta)}\right)^{Y_k}\,, \end{aligned} \quad (15)$$

where $\alpha(\theta) = \int_0^\infty (2\pi\gamma^2)^{-1/2}\exp\{-(1/2\gamma^2)(t-\theta)^2\}\mathrm{d}t$ and clearly belongs to the exponential family. The pdf of the posterior distribution π and any distribution $\tilde{\pi}_n(U) \in \mathsf{F}_n$ writes respectively as

$$\pi(\theta\,|\,Y_{1:N}) \propto p(\theta)\,(1-\alpha(\theta))^N\left(\frac{\alpha(\theta)}{1-\alpha(\theta)}\right)^{\sum_{k=1}^N Y_k}\,,$$

$$\tilde{\pi}_n(\theta\,|\,Y_U) \propto p(\theta)\,(1-\alpha(\theta))^N\left(\frac{\alpha(\theta)}{1-\alpha(\theta)}\right)^{(N/n)\sum_{k\in U} Y_k}\,,$$

where p is a prior density on θ. Again, in this example, the posterior density is easy to evaluate pointwise, even when N is extremely large, as it only requires to sum over all the binary variables $Y_1, \ldots, Y_N$. As a consequence, samples from π can routinely be obtained by a standard M–H algorithm and similarly for any distribution $\tilde{\pi}_n(U) \in \mathsf{F}_n$.

We simulated $N = 10{,}000$ simulated data $Y_1, \ldots, Y_N$ from (14), with true parameter $\theta^* = 1$. We used the prior distribution $p = \mathcal{N}(0, 10)$. In this probit model, S is simply the identity function, implying that $\|\Delta_n(U)\|$ gives the absolute value of the difference between the scaled proportion of 1 and 0's between the full dataset and the subset Y_U. Figure 2 reports the density functions of π and several other distributions $\tilde{\pi}_n(U) \in \mathsf{F}_n$, for $n = 100$ and $n = 1000$, with different values for the quantity $\|\Delta_n(U)\|$ (7). This plot, as well as the quantitative result of Table 1 are consistent with the statement of Corollary 1: when learning from a subsample of n data, one should work with a subset U featuring a perfect match with the full dataset, i.e $\|\Delta_n(U)\| = 0$, or as small as possible to achieve an *optimal* approximation of π. Finally, Fig. 3 illustrates Proposition 3: assigning the distribution $\nu_{n,\epsilon}$

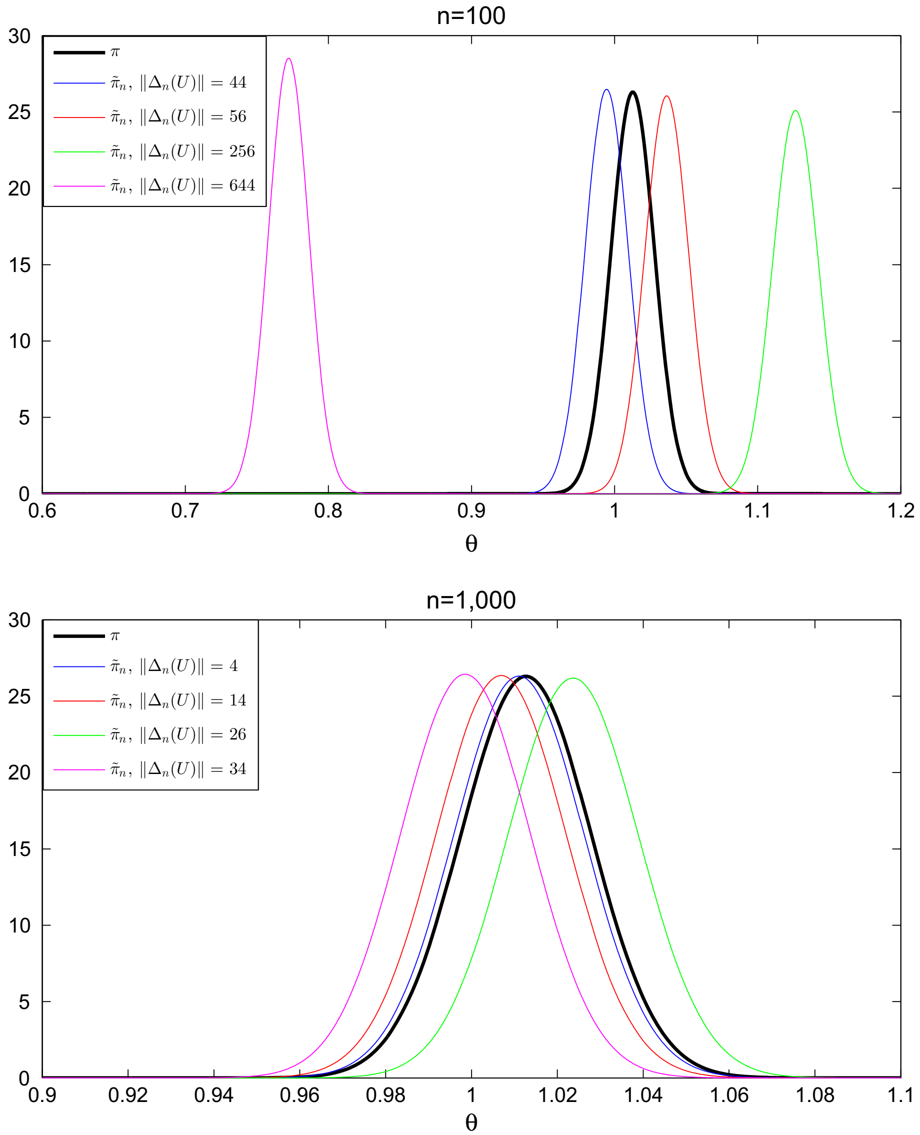


Fig. 2 (Example 2: Probit model) Influence of the parameter $U \in \mathsf{U}_n$ on the sub-posterior distribution $\tilde{\pi}_n(U)$ and comparison with π for subsets of size $n = 100$ (top) and $n = 1000$ (bottom)

(12) to the subsamples allows one to control the expectation of the likelihood ratio $f(Y \mid \theta)/f(Y_U \mid \theta)^{N/n}$ around 1.

4 Informed sub-sampling MCMC

In this section, we do not assume any particular correlation pattern for the sequence of observations, nor any specific likelihood model and simply write the posterior distribution π as

$$\pi(\mathrm{d}\theta \mid Y_{1:N}) \propto p(\mathrm{d}\theta) f(Y_{1:N} \mid \theta)\,. \tag{16}$$

The Informed Sub-Sampling MCMC (ISS-MCMC) methodology that we describe now can be regarded as an extension of the approximation detailed in the previous section to non-exponential family models with possibly dependent observations.

4.1 Motivation of our approach

Central to our approach is the idea that all subsamples Y_U ($U \in \mathsf{U}_n$) are not equally valuable for inferring π. Here, we do not assume the existence of a sufficient statistic mapping for the models under consideration. Thus, in order to discriminate between different subsamples, we introduce an *artificial* summary statistic mapping $S : \mathsf{Y}_n \to \mathsf{S}$ ($n \leq N$), where $\mathsf{S} \subseteq \mathbb{R}^s$. The choice of the summary statistics S is problem specific and is meant to be the counterpart of the sufficient statistic mapping for general models (hence sharing, slightly abusively, the same notation). Since the question of specifying summary statistics also arises in Approxi-

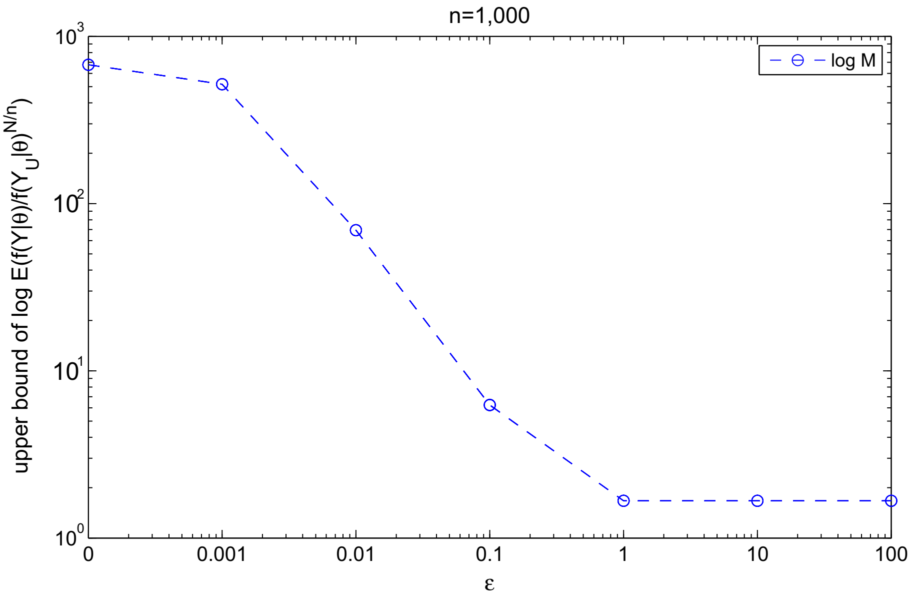


Fig. 3 (Example 2: Probit model) Influence of the parameter ϵ of the distribution $\nu_{n,\epsilon}$ on the upper bound M of $\mathbb{E}\{f(Y \mid \theta)/f(Y_U \mid \theta)^{N/n}\}$ for $\theta \in (0, 1.5)$, for $n = 1000$. When $\epsilon = 0$, $\nu_{n,\epsilon}$ is uniform (i.e. an identical weight is assigned to all the subsamples) and as a consequence $M \equiv \infty$. Conversely, when $\epsilon \gg 0$, the mass of $\nu_{n,\epsilon}$ spreads over the best subsamples Y_U, $U \in \mathsf{U}_n$ (i.e. those minimizing $\Delta_n(U)$) and the bound M is smaller than e^2. Indeed, by assigning a weight $\nu_{n,\epsilon}(U) \propto \exp\{-\epsilon\Delta_n(U)^2\}$ those subsamples Y_U, $U \in \mathsf{U}_n$ that have a large $\Delta_n(U)$ will yield a negligible contribution to the expectation, hence preventing from divergence

mate Bayesian Computation (ABC), one can take advantage of the abundant ABC literature on this topic to find some examples of summary statistics for usual likelihood models (see e.g. Nunes and Balding 2010; Csilléry et al. 2010; Marin et al. 2012; Fearnhead and Prangle 2012). More details on validation of summary statistics are discussed in Sect. 5.3.2.

Because the statistics used to assess the representativeness of a subsample Y_U w.r.t. the full dataset Y are only *summary* and not *sufficient*, the results of Sect. 3 are no longer valid. In particular, should an optimal subset U^* minimising a distance between $S(Y_U)$ and $S(Y)$ exist, inferring π through the approximation $\tilde{\pi}_n(U^*)$ is in no sense optimal. In fact, as shown in several examples of Sect. 6, this approximation is usually poor. In such a setting, it is reasonable to consider extending the set of subsamples of interest to a pool of *good* subsamples. This naturally suggests using the distribution $\nu_{n,\epsilon}$ (12) to discriminate between the subsamples, replacing sufficient by summary statistics and relaxing the assumption $\epsilon \to \infty$, in order to account for a collection of good subsamples. Before proceeding to the presentation of our algorithm, we define the following quantities related to a subset $U \in \mathsf{U}_n$:

$$\bar{S}(Y_U) = S(Y_U)/n\,, \qquad \bar{\Delta}_n(U) = S(Y)/N - S(Y_u)/n\,. \tag{17}$$

4.2 Informed sub-sampling MCMC: the methodology

Informed Sub-Sampling MCMC is a scalable adaptation of the Metropolis–Hastings algorithm (Metropolis et al. 1953), designed for situations when N is prohibitively large to perform inference on the posterior π in a reasonable time frame. ISS-MCMC relies on a Markov chain whose transition kernel has a bounded computational complexity, which can be controlled through the parameter n. We first recall how the Metropolis–Hastings algorithm produces a π-reversible Markov chain $\{\theta_i,\ i \in \mathbb{N}\}$, for any distribution π known up to a normalizing constant. The index i is used hereafter to refer to the Markov chain iteration.

4.2.1 Metropolis–Hastings

Let Q be a transition kernel on $(\Theta, \mathcal{B}(\Theta))$ and assume that the Metropolis–Hastings Markov chain is at state θ_i. A transition $\theta_i \to \theta_{i+1}$ consists in the two following step:

(a) propose a new parameter $\theta \sim Q(\theta_i, \cdot)$
(b) set the next state of the Markov chain as $\theta_{i+1} = \theta$ with probability

$$\alpha(\theta_i, \theta) = 1 \wedge a(\theta_i, \theta)\,, \; a(\theta_i, \theta) = \frac{\pi(\theta \mid Y)Q(\theta, \theta_i)}{\pi(\theta_i \mid Y)Q(\theta_i, \theta)} \tag{18}$$

and as $\theta_{i+1} = \theta_i$ with probability $1 - \alpha(\theta_i, \theta)$.

Algorithm 1 details how to simulate a Metropolis–Hastings Markov chain $\{\theta_i,\ i \in \mathbb{N}\}$.

Algorithm 1 Metropolis–Hastings algorithm

```
Input: initial state $\theta_0$ and posterior evaluation $\pi(\theta_0 \mid Y)$
for $i = 1, 2, \ldots$ do
    propose a new parameter $\theta \sim Q(\theta_{i-1};\ \cdot)$ and draw $I \sim$ unif(0, 1)
    compute $\pi(\theta \mid Y)$ and $a = a(\theta_{i-1}, \theta)$ defined in (18)
    if $I \leq a$ then
        set $\theta_i = \theta$
    else
        set $\theta_i = \theta_{i-1}$
    end if
end for
return: the Markov chain $\{\theta_i,\ i \in \mathbb{N}\}$
```

4.2.2 Informed sub-sampling MCMC

To avoid any confusion, we denote by $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ the sequence of parameters generated by the Informed Sub-Sampling Markov chain, by contrast to the Markov chain $\{\theta_i,\ i \in \mathbb{N}\}$ produced by the Metropolis–Hastings algorithm (Algorithm 1). The pool of *good* subsamples used in the Informed Sub-Sampling inference is treated as a sequence of missing data $U_1, U_2, \ldots$ and is thus simulated by our algorithm. More precisely, ISS-MCMC produces a Markov chain $\{(\tilde{\theta}_i, U_i),\ i \in \mathbb{N}\}$ on the extended space $\Theta \times \mathsf{U}_n$. Inspired by the analysis of Sect. 3, the sequence of subsamples $\{U_i,\ i \in \mathbb{N}\}$ is randomly updated in a way that favours those subsets whose summary statistics vector is close to that of the full dataset. Let R be a symmetric transition kernel on $(\mathsf{U}_n, \mathcal{U}_n)$, a transition $(\tilde{\theta}_i, U_i) \to (\tilde{\theta}_{i+1}, U_{i+1})$ consists in the two following steps:

(i) (a) propose a new subset variable $U \sim R(U_i,\ \cdot)$
(b) set $U_{i+1} = U$ with probability

$$\beta(U_i, U) = 1 \wedge b(U_i, U)\,,$$
$$b(U_i, U) = \exp\left\{\epsilon\left(\|\Delta_n(U_i)\|^2 - \|\Delta_n(U)\|^2\right)\right\} \tag{19}$$

and $U_{i+1} = U_i$ with probability $1 - \beta(U_i, U)$. Δ_n is defined at Eq. (7).

(ii) (a) propose a new parameter $\tilde{\theta} \sim Q(\tilde{\theta}_i, \cdot)$
(b) set $\tilde{\theta}_{i+1} = \tilde{\theta}$ with probability

$$\tilde{\alpha}(\tilde{\theta}_i, \tilde{\theta}) = 1 \wedge \tilde{a}(\tilde{\theta}_i, \theta \mid U_{i+1})\,,$$
$$\tilde{a}(\tilde{\theta}_i, \tilde{\theta} \mid U_{i+1}) = \frac{\tilde{\pi}_n(\tilde{\theta} \mid Y_{U_{i+1}})Q(\tilde{\theta}, \tilde{\theta}_i)}{\tilde{\pi}_n(\tilde{\theta}_i \mid Y_{U_{i+1}})Q(\tilde{\theta}_i, \tilde{\theta})} \tag{20}$$

and as $\tilde{\theta}_{i+1} = \tilde{\theta}_i$ with probability $1 - \tilde{\alpha}(\tilde{\theta}_i, \tilde{\theta} \mid U_{i+1})$.

Algorithm 2 details how to simulate an Informed Sub-Sampling Markov chain. Note that at step 11, if $U_i = U_{i-1}$, the quantity $\tilde{\pi}_n(\theta_{i-1} \mid U_i)$ has already been calculated at the previous iteration.

Algorithm 2 Informed Sub-Sampling MCMC algorithm

```
1: Input: initial state $(\tilde{\theta}_0, U_0)$ and summary statistics $S_0 = \bar{S}(Y_{U_0})$, $S^* = \bar{S}(Y)$
2: for $i = 1, 2, \ldots$ do
3:     propose a new subset $U \sim R(U_{i-1}, \cdot)$ and draw $J \sim$ unif(0, 1),
4:     compute $S = \bar{S}(Y_U)$ and $b = b(U_{i-1}, U)$ defined in (19)
5:     if $J \leq b$ then
6:         set $U_i = U$ and $S_i = S$
7:     else
8:         set $U_i = U_{i-1}$ and $S_i = S_{i-1}$
9:     end if
10:    propose a new parameter $\tilde{\theta} \sim Q(\tilde{\theta}_{i-1};\ \cdot)$ and draw $I \sim$ unif(0, 1)
11:    compute $\tilde{\pi}_n(\tilde{\theta}_{i-1} \mid Y_{U_i})$, $\tilde{\pi}_n(\tilde{\theta} \mid Y_{U_i})$ and $\tilde{a} = \tilde{a}(\tilde{\theta}_{i-1}, \tilde{\theta} \mid U_i)$ defined in (20)
12:    if $I \leq \tilde{a}$ then
13:        set $\tilde{\theta}_i = \tilde{\theta}$
14:    else
15:        set $\tilde{\theta}_i = \tilde{\theta}_{i-1}$
16:    end if
17: end for
18: return: the Markov chain $\{(\tilde{\theta}_i, U_i),\ i \in \mathbb{N}\}$
```

4.3 Connection with noisy ABC

Approximate Bayesian Computation (ABC) is a class of statistical methods, initiated in Pritchard et al. (1999), that allows one to infer π in situations where the likelihood f is intractable but forward simulation of pseudo data $Z \sim f(\cdot \mid \theta)$ is doable. More precisely, the algorithm consisting of (i) $\vartheta \sim p$, (ii) $Z \sim f(\cdot \mid \vartheta)$ and (iii) set $\theta = \vartheta$ only if $\{Z = Y\}$, does produce a sample θ whose distribution is $\pi(\cdot \mid Y)$. Regarding the situation $N \to \infty$ as a source of intractability, one could attempt to borrow from ABC to sample from π. However, since $N \to \infty$, sampling from the likelihood model is impossible and a natural idea is to replace step (ii) by drawing subsamples Y_U ($U \in \mathsf{U}_n$), leading to what we refer as Informed Sub-Sampling, as opposed to Informed Sub-Sampling MCMC described in the previous Subsection. Obviously, the event $\{Y_U = Y\}$ is impossible except in the trivial situation where $N = n$. Overcoming situations where $\{Y = Z\}$ is impossible or very unlikely has already been addressed in the ABC literature [see Fearnhead and Prangle (2012) and Wilkinson (2013)], leading to approximate ABC algorithms. In particular, step (iii) is replaced by a step that sets $\theta = \vartheta$ with probability $\propto \exp\{-\epsilon\|S(Z) - S(Y)\|^2\}$ where S is a vector of summary statistics and $\epsilon > 0$ a

Table 2 Comparison between ABC and Informed Sub-Sampling, an adaptation of ABC designed for situations where $N \gg 1$ and likelihood simulation is not possible

Step	ABC		Informed Sub-Sampling	
(i)	$\vartheta \sim p$		–	
(ii)	$Z \sim f(\,\cdot \mid \vartheta)$		$Z = Y_U$, $U \sim \mathrm{unif}(\mathsf{U})$	
(iii)	Exact	Noisy	Exact	Noisy
	if $Z = Y$,	with proba. $\propto$	if $Z = Y$	with proba. $\propto$
	set $\theta = \vartheta$	$e^{-\epsilon\\|S(Y)-S(Z)\\|^2}$	draw $\theta \sim \pi(U)$	$e^{-\epsilon\\|S(Y)-(N/n)S(Y_U)\\|^2}$
		set $\theta = \vartheta$		draw $\theta \sim \pi(U)$

The exact algorithms provide samples from π while the noisy algorithms sample from approximation of π given in (21) and (22)

tolerance parameter. We build on this analogy to propose a noisy Informed Sub-Sampling algorithm, see Table 2 for more details.

The Noisy ABC algorithm replaces direct inference of π by the following surrogate distribution

$$\begin{aligned}\hat{\pi}_{\mathrm{ABC}}(\mathrm{d}\theta \mid Y) &:\propto p(\mathrm{d}\theta)\hat{f}_{\mathrm{ABC}}(Y \mid \theta)\\ &= p(\mathrm{d}\theta)\int f(\mathrm{d}Z \mid \theta)\exp\{-\epsilon\|S(Z)-S(Y)\|^2\}\,, \qquad (21)\end{aligned}$$

where the exact likelihood is replaced by $\hat{f}_{\mathrm{ABC}}$. Similarly, the approximation of π stemming from Informed Sub-Sampling is:

$$\begin{aligned}\hat{\pi}_n(\mathrm{d}\theta \mid Y) &:\propto p(\mathrm{d}\theta)\hat{f}(Y \mid \theta)\\ &= p(\mathrm{d}\theta)\sum_{U \in \mathsf{U}_n} f^{(N/n)}(Y_U \mid \theta)\\ &\quad \exp\{-\epsilon\|(N/n)S(Y_U)-S(Y)\|^2\}\,. \qquad (22)\end{aligned}$$

This analogy shows that there is a connection between the ABC and the Informed Sub-Sampling in the way both approximate π, see (21) and (22). However, since sampling from $\nu_{n,\epsilon}$ and $\pi_n(U)$ are not feasible, this approach cannot be considered, hence motivating the use of Markov chains instead, i.e Informed Sub-Sampling MCMC. Moreover, quantifying the approximation of π by $\hat{\pi}_n$ (22) is technically challenging while resorting to the Informed Sub-Samping Markov chain allows to use the Noisy MCMC framework developed in Alquier et al. (2016) to quantify this approximation. This is the purpose of the following Section.

5 Theoretical analysis of informed sub-sampling MCMC

By construction, ISS-MCMC samples a Markov chain on an extended state space $\{(\tilde{\theta}_i, U_i),\ i \in \mathbb{N}\}$ but the only useful outcome of the algorithm for inferring π is the marginal chain $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$. In this section, we study the distribution of the marginal chain and denote by $\tilde{\pi}_i$ the distribution of the random variable $\tilde{\theta}_i$ for some iteration $i \in \mathbb{N}$. Note that $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ is identical to the Metropolis–Hastings chain $\{\theta_i,\ i \in \mathbb{N}\}$, up to replacing α by $\tilde{\alpha}$ in the accept/reject step. This change, from which the computational gain of our method originates, has important consequences on the stability of the Markov chain and, in particular, implies that π is not the stationary distribution of $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$. Interest lies in quantifying the distance between $\tilde{\pi}_i$ and π. In this paper, our results are expressed in total variation distance but the recent works of Rudolf and Schweizer (2018) and Johndrow and Mattingly (2017) suggest that carrying out the analysis using the Wasserstein metric may lead to more accurate bounds when Θ is not compact. We first recall the definition of the total variation distance which, for two distributions with density function π and $\tilde{\pi}_i$ respectively w.r.t. the same common dominating measure, denoted $\mathrm{d}\theta$, can be expressed as

$$\|\pi - \tilde{\pi}_i\| = (1/2)\int_{\Theta} |\pi(\theta) - \tilde{\pi}_i(\theta)|\mathrm{d}\theta\,.$$

5.1 Assumptions

Let K denote the *exact* M–H transition kernel, with proposal Q, described in Algorithm 1. Q is fixed and set as a random walk kernel that achieves a reasonable acceptance rate. By construction, K is π-reversible and thus π-invariant. Moreover, K is assumed to be ergodic i.e $\|K(x,\ \cdot) - \pi\| \to 0$ at a geometric rate and the convergence is either simple (Assumption **A**.1) or uniform (Assumption **A**.2).

A 1 Geometric ergodicity There exists a constant $\varrho \in (0, 1)$ and a function $C : \Theta \to \mathbb{R}^+$ such that for all $(\theta_0, i) \in \Theta \times \mathbb{N}$

$$\|\pi - K^i(\theta_0, \cdot)\| \le C(\theta_0)\varrho^i\,. \qquad (23)$$

A 2 Uniform ergodicity There exists two constants $C < \infty$ and $\varrho \in (0, 1)$ such that for all $i \in \mathbb{N}$

$$\sup_{\theta_0 \in \Theta} \|\pi - K^i(\theta_0,\ \cdot)\| \le C\varrho^i\,. \qquad (24)$$

As observed in Remark 2 (Appendix A.6), the ISS-MCMC marginal Markov chain $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ is time inhomogeneous.

Indeed, conditionally on $\tilde{\theta}_i$, the probability of the transition $\tilde{\theta}_i \to \tilde{\theta}_{i+1}$ depends on the iteration index i. This complicates the analysis of ISS-MCMC as most results on perturbation of Markov chains are established for time homogeneous Markov chains. For simplicity, we present in this section an analysis of a slight variation of ISS-MCMC that assumes independence between the different subsets $\{U_i,\ i \in \mathbb{N}\}$ (Assumption **A.**3).

A 3 **IID subsets** The subsets $U_1, U_2, \ldots$ are independent and identically distributed under $\nu_{n,\epsilon}$.

In practice, Assumption **A.**3 is satisfied when steps (3)-(9) of Algorithm 2 are repeated a large number of times to simulate U_{i+1} given U_i. Under **A.**3, $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ is a time homogeneous Markov chain whose transition kernel $\tilde{K}_{n,\epsilon}$ is

$$\forall\,(\tilde{\theta}, A) \in \Theta \times \vartheta, \qquad \tilde{K}_{n,\epsilon}(\tilde{\theta}, A) = \sum_{u \in \mathsf{U}_n} K(\tilde{\theta}, A \mid u)\nu_{n,\epsilon}(u)\,, \tag{25}$$

where for all $\theta \in \Theta$, defined at Eq. (5) $K(\theta, \cdot \mid u)$ is the *exact* M–H transition kernel conditionally on some $\theta \in \Theta$ with proposal Q that targets $\tilde{\pi}_n(\cdot \mid Y_u)$ (5).

The results stated in this section hold under Assumption **A.**3. We nevertheless note that this assumption might be relaxed. In particular, we show how the analysis carried out under uniform ergodicity assumption can be extended even if Assumption **A.**3 does not hold, see Appendix A.6. In the geometric ergodic case, a similar extension may be doable, see e.g. (Douc et al. 2004, Theorem 8), but this is out of the scope of this paper. In general, the perturbation bounds with time inhomogeneous kernels are more obscure to interpret. Note that the numerical illustrations of ISS-MCMC presented at Sect. 6 were performed without satisfying Assumption **A.**3, i.e implementing Algorithm 2, and lead to satisfactory results.

Finally, we consider the following assumption for the summary statistics mapping. This assumption is motivated at two levels. First, it is necessary to have some assumptions on the summary statistics to derive theoretical results for ISS-MCMC in absence of sufficient statistics. Second, it offers a way to validate empirically the choice of summary statistics for a given model, see Sect. 6.

A 4 **Summary Statistics** There exists a constant $\gamma_n < \infty$, such that for all $(\theta, U) \in \Theta \times \mathsf{U}_n$

$$\begin{aligned} &|\log f(Y \mid \theta) - (N/n)\log f(Y_U \mid \theta)| \\ &\quad \le \gamma_n N \|\bar{S}(Y) - \bar{S}(Y_U)\|\,. \end{aligned} \tag{26}$$

Assumption **A.**4 imposes a condition simultaneously on the model f and the summary statistics S. In particular, it assumes that for any $\theta \in \Theta$, the variation of the scaled likelihood of the subsamples Y_U $(U \in \mathsf{U}_n)$ around $f(Y \mid \theta)$ is controlled by the distance between the full dataset Y and the subsample Y_U, as measured through their summary statistics. This is a strong assumption which is unlikely to hold if Θ is not a compact set. It implies that even in absence of sufficient statistics, a result similar to Proposition 3 exists. One can also note that when $n \to N$, the constant γ_n in Eq. (26) goes to zero.

5.2 *K* is geometrically ergodic

Our main result is that for a sufficiently large size of subsample n, ISS-MCMC admits a stationary distribution. This follows from an adaptation of the work of Medina-Aguayo et al. (2016) to the context of ISS-MCMC.

Proposition 4 *Assume that Assumptions* **A.**1 *and* **A.**3 *and* **A.**4 *hold, then there exists an* $n_0 \le N$ *such that for all* $n > n_0$, $\tilde{K}_{n,\epsilon}$ *is also geometrically ergodic for all* $\epsilon > 0$.

The proof is outlined to Appendix A.4.

A direct consequence of Proposition 4, see for instance (Meyn and Tweedie 2009, Theorem 16.0.1), is that for n sufficiently large, $\tilde{K}_{n,\epsilon}$ admits a stationary distribution and that this stationary distribution converges to π as $n \to N$. In most cases, it is difficult to obtain the rate of convergence of γ_n (26) to 0 under the assumption that K is geometrically ergodic. We nevertheless note that this rate is related to rate of convergence of γ_n to 0 as hinted by Medina-Aguayo et al. (2016, Theorem 4.1).

5.3 *K* is uniformly ergodic

In addition to admitting a stationary distribution for a large enough n, we now show that under the assumption of uniform ergodicity it is possible to quantify the rate of convergence. Our main result follows from an adaptation of the work of Alquier et al. (2016) to the context of ISS-MCMC.

Proposition 5 *Define*

$$A_n := \mathbb{E}\left\{\sup_{\theta \in \Theta} \frac{1}{\phi_U(\theta)}\right\} = \sum_{U \in \mathsf{U}_n} \nu_{n,\epsilon}(U) \sup_{\theta \in \Theta} \frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}}\,, \tag{27}$$

where for all $(\theta, U) \in (\Theta \times \mathsf{U}_n)$, *we have set* $\phi_U(\theta) := f(Y_U \mid \theta)^{N/n} / f(Y \mid \theta)$ *and*

$$\begin{aligned} B_n(\theta, U) &:= \mathbb{E}\{a(\theta, \theta')|\phi_U(\theta) - \phi_U(\theta')|\} \\ &= \int Q(\theta, d\theta')a(\theta, \theta')|\phi_U(\theta) - \phi_U(\theta')|\,. \end{aligned} \tag{28}$$

Assume that Assumptions **A.**2*,* **A.**3 *and* **A.**4 *hold, then there exists a constant* $\kappa < \infty$ *such that for all* $i \in \mathbb{N}$

$$\|K^i(\theta_0,\,\cdot) - \tilde{K}^i_{n,\epsilon}(\theta_0,\,\cdot)\| \le \kappa A_n \sup_{(\theta,U)\in\Theta\times\mathsf{U}_n} B_n(\theta,U)\,, \tag{29}$$

and

$$\lim_{i\to\infty}\sup_{\theta\in\Theta}\|\pi - \tilde{K}^i_{n,\epsilon}(\theta,\,\cdot)\| = \kappa A_n \sup_{(\theta,U)\in\Theta\times\mathsf{U}_n} B_n(\theta,U)\,. \tag{30}$$

Moreover, for a large enough subset size n, the marginal Markov chain produced by ISS-MCMC admits an invariant distribution $\tilde{\pi}_n$ *that satisfies*

$$\|\pi - \tilde{\pi}_n\| \le \kappa A_n \sup_{(\theta,U)\in\Theta\times\mathsf{U}_n} B_n(\theta,U)\,. \tag{31}$$

The proof of Proposition 5 is postponed to Appendix A.5. Note that an extension of this result to the case where Assumption **A.**3 does not hold is presented at Appendix A.6.

Since for any two measures (μ,μ'), $\|\mu-\mu'\| \le 1$, the upper bounds of Proposition 5 are only informative if there are smaller than 1. Those bounds are a product of two expectations. We now show how those two expectations are controlled respectively through the choice of proposal kernel and the choice of summary statistics.

5.3.1 Choice of the proposal kernel

Assuming a Gaussian random walk proposal with covariance matrix $\Sigma^T\Sigma$, B_n can be expressed as $B_n(\theta) = \sup_{U\in\mathsf{U}_n} D_1(U,\theta)$ where D_1 is defined as

$$D_1(U,\theta) := \int \Phi_d(\mathrm{d}\zeta)\frac{\pi(\theta+\Sigma\zeta)}{\pi(\theta)}|\phi_U(\theta) - \phi_U(\theta+\Sigma\zeta)|\,, \tag{32}$$

where Φ_d is the standard Gaussian distribution in dimension $d = \dim(\Theta)$. When $N \gg 1$, the Bernstein-von Mises theorem states that, under conditions on the likelihood function, the posterior distribution can be approximated by a Gaussian with mean set as the maximum likelihood estimator θ^* and covariance $I(\theta_0)^{-1}/N$ where I is the Fisher information matrix and $\theta_0 \in \Theta$ some parameter. Since ISS-MCMC aims at sampling from an approximation of π, setting $\Sigma = (1/\sqrt{N})M$ where M^TM is an approximation of $I(\theta_0)^{-1}$ is a reasonable choice. Proposition 6 shows that D_1 can, in this scenario, be arbitrarily brought down close to 0.

Proposition 6 *Under the assumption that the proposal kernel* Q *is a Gaussian Random Walk with covariance matrix* $\Sigma = (1/\sqrt{N})M$*, we have*

$$D_1(U,\theta) \le \frac{\|\nabla_\theta\phi_U(\theta)\|}{\sqrt{N}}\left\{\sqrt{\frac{2}{\pi}}\|M\|_1 + \frac{\|M\|_2^2\|\nabla_\theta\log\pi(\theta)\|}{\sqrt{N}}\right\} + \frac{d}{2N}|\!|\!|M^T\nabla^2_\theta\phi_U(\theta)M|\!|\!| + \mathbb{E}\{R(\|M\zeta\|/\sqrt{N})\}\,, \tag{33}$$

where $R(x) =_{x\to 0} o(x)$ *and for any square matrix* M *of dimension* $\mathbb{R}^d$*, we have set* $\|M\|_1 := \sum_{1\le i,j\le d}|M_{i,j}|$*,* $\|M\|_2 := \{\sum_{1\le i,j\le d} M_{i,j}^2\}^{1/2}$ *and* $|\!|\!|\cdot|\!|\!|$ *is the operator norm.*

The proof is postponed to Appendix A.7. Under regularity assumptions on the likelihood model, the gradient of $\log\pi$ and ϕ_U and the Hessian of ϕ_U are bounded and the upper bound of $D_1(U,\theta)$ can be brought down arbitrarily to 0, uniformly in (U,θ), through M when $N \gg 1$.

5.3.2 Choice of the summary statistics

In Eq. (27), the likelihood of each subsample is raised at the power N/n (i.e typically several orders of magnitude) and therefore subsamples unlikely under $f(\cdot\,|\,\theta)$ will contribute to make A_n very large. Ideally the choice of S would guarantee that subsamples Y_U having a very small likelihood $f(Y_U\,|\,\theta)$ are assigned to a weight $\nu_{n,\epsilon}(U) \approx 0$ to limit their contribution. In other words, S should be specified in a way that prevents $f(Y_U\,|\,\theta)$ to go to 0 at a rate faster than $\nu_{n,\epsilon}(U)$. This is ensured if Assumption **A.**4 holds. Indeed, in such a case

$$\begin{aligned}
A_n &= \sum_{U\in\mathsf{E}_n(\theta)} \nu_{n,\epsilon}(U)\sup_{\theta\in\Theta}\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}} \\
&\quad + \sum_{U\in\mathsf{U}_n\backslash\mathsf{E}_n(\theta)} \nu_{n,\epsilon}(U)\sup_{\theta\in\Theta}\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}} \\
&\le \nu_{n,\epsilon}\left(\mathsf{E}_n(\theta)\right) + \sum_{U\in\mathsf{U}_n\backslash\mathsf{E}_n(\theta)} \exp\{-\epsilon\|\Delta_n(U)\|^2 \\
&\quad + \gamma_n\|\Delta_n(U)\| - \log Z_n(\epsilon)\}\,,
\end{aligned}$$

where we have defined $\mathsf{E}_n(\theta) := \{U \in \mathsf{U}_n,\ \sup_{\theta\in\Theta} f(Y\,|\,\theta)/f(Y_U\,|\,\theta)^{N/n} < 1\}$ and $Z_n(\epsilon) = \sum_{U\in\mathsf{U}_n}\exp\{-\epsilon\|\Delta_n(U)\|^2\}$. Clearly, if ϵ has the same order of magnitude as γ_n, each term of the sum remains bounded when $\|\Delta_n(U)\| \to \infty$. Conversely, setting $\epsilon = 0$ is equivalent to choosing $\nu_{n,\epsilon}$ as the uniform distribution on U_n and may not allow to bound A_n, see Fig. 3 related to the probit example.

Potential summary statistics can be empirically validated by checking that they satisfy Assumption **A.**4. This validation can be performed graphically, by repeating the following operations for a number of parameters $\theta_k \sim_{\text{i.i.d}} p$:

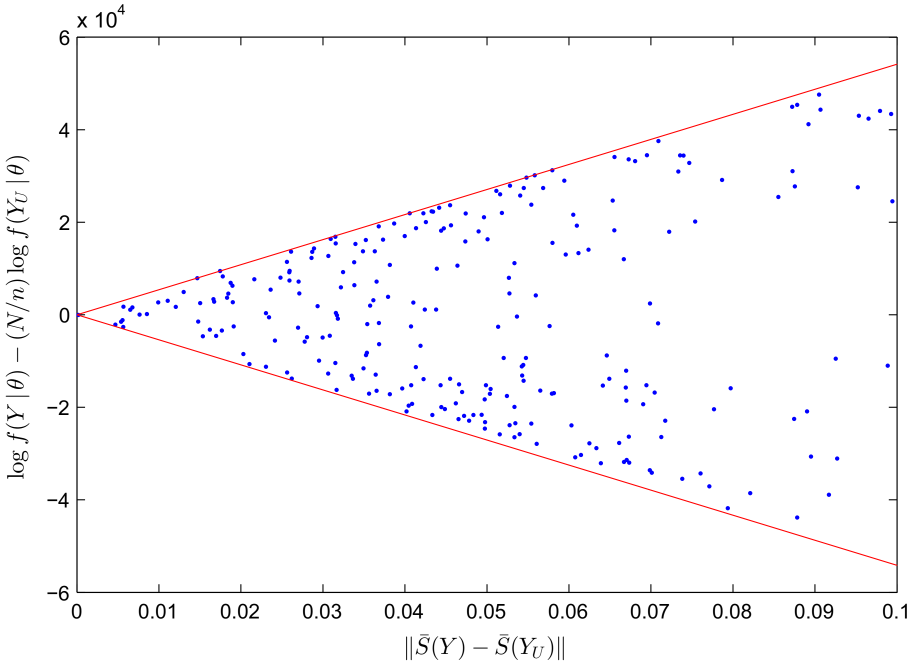


Fig. 4 (Example 3: Autoregressive time series) Validation of summary statistics set as the solution of Yule-Walker equation, with $n = 1000$. This choice of sufficient statistics satisfies Assumption **A.**4 with $\gamma_n \approx 5\,10^6$. Each dot corresponds to a point $(\log f(Y\mid\theta) - (N/n)\log f(Y_U\mid\theta), \|\bar{\Delta}_n(U)\|)$ where θ and U were respectively drawn from the prior p and uniformly at random in U_n. The red lines allow to estimate the parameter γ_n of Assumption **A.**4

(i) draw subsets $U_1, U_2, \ldots$ uniformly at random in U_n,
(ii) plot the points with coordinates

$$(x_{k,i}, y_{k,i}) = (\|\Delta_n(U_i)\|, \log f(Y\mid\theta_k) \\ -(N/n)\log f(Y_{U_i}\mid\theta_k))\,.$$

The statistics are validated if there exists $\gamma_n < \infty$ such that the points $(x_{k,i}, y_{k,i})$ satisfy $|y_{k,i}/x_{k,i}| \leq \gamma_n$, as illustrated at Fig. 4.

In situations where the maximum likelihood estimator $\theta^*(Y_{1:n})$ is easy and quick to evaluate numerically, we recommend setting $\bar{S}(Y_{1:n}) = \theta^*(Y_{1:n})$. In the case of independent observations of a well-specified model, setting the summary statistics as the maximum likelihood estimate is justified by the following Proposition which implies that Assumption **A.**4 holds, asymptotically, up to a constant.

Proposition 7 *We assume that the whole dataset comprises $N = \rho n$ independent observations and there exists some $\theta_0 \in \Theta$ such that $Y_i \sim f(\,\cdot\mid\theta_0)$. Let θ^* be the MLE of $Y_1, \ldots, Y_N$ and θ^*_U be the MLE of the subsample Y_U ($U \in \mathsf{U}_n$). Then, there exists a constant β, a metric $\|\cdot\|_{\theta_0}$ on Θ and a non-decreasing subsequence $\{\sigma_n\}_{n\in\mathbb{N}}$, ($\sigma_n \in \mathbb{N}$) such that for all $U \subset \{1, 2, \ldots, \rho\sigma_n\}$ with $|U| = \sigma_n$, we have for p-almost all θ in an neighborhood of θ_0:*

$$\log f(Y_{1:\rho\sigma_n}\mid\theta) - \rho\log f(Y_U\mid\theta) \leq H_n(Y,\theta) + \beta \\ + \frac{\rho n}{2}\|\theta^*_U - \theta^*\|_{\theta_0}\,, \tag{34}$$

where

$$\operatorname*{plim}_{n\to\infty} \quad H_n(Y,\theta) \overset{\mathbb{P}_{\theta_0}}{=} 0\,.$$

The proof is detailed in Appendix B and follows from a careful application of a Bernstein-von Mises theorem. Note that an extension of Proposition 7 to cases where the observations are not independent may exist provided that a Bernstein-von Mises theorem holds for the model at hand, which is the case for dependent observations if the likelihood model satisfies local asymptotic normality conditions Le Cam (1953, 1986).

Note that since Assumption **A.**4 is mostly used in Propositions 4 and 5 to guarantee that the log-likelihood ratio between the likelihood and the scaled likelihood of a subsample is bounded, the constant β in Proposition 7 is not a major concern. In addition, it is straightforward to see that this constant vanishes when the subset size grows faster than the full dataset, i.e $\rho \downarrow 1$, once in the asymptotic regime of Eq. (34).

We remark that Proposition 7 is in line with the results regarding optimal summary statistics for ABC established in Fearnhead and Prangle (2012). The authors show that the quadratic error loss between the ABC estimate based on $\tilde{\pi}_{\text{ABC}}$ (Eq. 21) and the true parameter is minimized when setting the summary statistics as the posterior mean, a choice which asymptotically coincides with the maximum likelihood estimator.

Finally, we note that, similarly to any approximate MCMC method, ISS-MCMC does not guarantee a Law of Large number for π-integrable functionals. However, assuming that the

M–H chain K is geometrically ergodic, it is straightforward to establish that for a large enough n,

$$\lim_{i\to\infty}\left|\frac{1}{i}\sum_{j=1}^{i} f(\tilde{\theta}_j) - \pi f\right| \leq 2\|f\|\|\pi - \tilde{\pi}_{n,\epsilon}\| \qquad \text{a.s}\,, \quad (35)$$

where $\pi f := \int f \mathrm{d}\pi$ and $\|f\| = \sup_{\theta\in\Theta}|f(\theta)|$. If in addition, K is uniformly ergodic Proposition 5 can help to bound the asymptotic error that typically arises in MCMC estimation of πf.

6 Illustrations

We evaluate the efficiency of ISS-MCMC on three different applications: inferring a time series observed at $N = 10^6$ contiguous time steps, a logistic regression with $N = 10^6$ observations and a Gaussian binary classification problem based on $N = 10^7$ data.

6.1 Implementation details of informed sub-sampling MCMC

Before illustrating the ISS-MCMC algorithm on the different examples, we address a few technical implementation details.

- On the subset size n: this parameter is essentially related to the computational budget available to the user. In the following example, we find out that using $n = N^{1/2}$ achieves a substantial computational gain at a price of a negligible bias.
- On the sufficient statistics S: to reduce the bias resulting from the Metropolis–Hastings approximation, S should be constructed so that Assumption **A.**4 holds. If the maximum likelihood estimator $\theta^*(Y)$ is quick to compute then Proposition 7 suggests that setting $S(Y) = \theta^*(Y)$ will theoretically satisfy Assumption **A.**4. Other sufficient statistics mapping can be used, typically those arising in the Approximate Bayesian Computation literature. In any case, we recommend checking Assumption **A.**4 graphically (see Sect. 5).
- On the bandwidth parameter ϵ: the theory shows that when $\epsilon \approx \gamma_n$, the asymptotic bias is controlled (γ_n is the constant in Assumption **A.**4). In practice, this may prove to be too large and could potentially cause the algorithm to get stuck on a very small number of subsets. To avoid such a situation, we suggest monitoring the refresh rate of subsamples that should occur with probability of at least 1%.
- On the initial subset U_0: in theory, one would run a preliminary Markov chain $\{U_1^{(0)}, \ldots, U_L^{(0)}\}$ (for some $L > 0$) targeting $\nu_{n,\epsilon}$, and set $U_0 = U_L^{(0)}$ in order for the results of Sect. 5.3.2 to hold. In practice, a more efficient approach is to use a simulating annealing Metropolis–Hastings algorithm, see Geyer and Thompson (1995). It introduces a sequence of tempered distributions $\nu_k := \nu_{n,\epsilon_k}$, such that $\epsilon_k = t_k\epsilon$ ($k \in \{1, \ldots, L\}$) where $t_1 = 0$ and $t_L = 1$. The transition kernel of the k-th iteration of the preliminary Markov chain is designed to be ν_k invariant. This technique facilitates sampling from a proxy of $\nu_{n,\epsilon}$ in a relative short time period as the successive tempered distributions help identifying those subsamples belonging to the high probability sets of $\nu_{n,\epsilon}$.

6.2 Inference of an AR(2) model

Example 3 An autoregressive time series of order 2 AR(2) $\{Y_k,\ k \leq N\}$ is defined recursively by:

$$\begin{cases} (Y_0, Y_1) \sim \mu := \mathcal{N}_2(\mathbf{0}_2, \theta_3^2\,\mathrm{Id}_2) \\ Y_n = \theta_1 Y_{n-1} + \theta_2 Y_{n-2} + Z_n\,, \quad Z_n \sim \mathcal{N}(0, \theta_3^2)\,, \qquad \forall\, n \geq 2, \end{cases} \quad (36)$$

where $\theta \in \Theta \subset \mathbb{R}^3$. The likelihood of an observed time series for this model writes

$$f(Y_{0:N} \mid \theta) = \mu(Y_{0:1}) \prod_{k=2}^{N} g(Y_k \mid Y_{k-1}, Y_{k-2}, \theta)\,, \quad (37)$$

such that for all $k \geq 1$,

$$g(Y_k \mid Y_{0:k-1}, \theta) = \Phi_1(Y_k;\ \theta_1 Y_{k-1} + \theta_2 Y_{k-2}, \theta_3^2) \quad (38)$$

where $x \to \Phi_1(x;\ m, v)$ is the pdf of the univariate Gaussian distribution with mean m and variance v.

This model has been used in Chib and Greenberg (1995, Section 7.2) to showcase the Metropolis–Hastings (M–H) algorithm. We follow the same setup here and in particular we use the same true parameter $\theta^* = (1, -.5, 1)$, same prior distribution and proposal kernel Q ; see Chib and Greenberg (1995) for more details. We sampled a time series $\{Y_k,\ k \leq N\}$ according to (36), with $N = 10^6$ under θ^*. Of course, in such a setup, M–H is prohibitively slow to be used in practice to sample from π as it involves evaluating the likelihood of the whole time series at each iteration. We nevertheless use M–H to obtain a ground truth of π.

For simplicity, we restrict the set of subsamples to n contiguous observations:

$$\{Y_{0:n-1}, Y_{1:n}, \ldots, Y_{N-n+1:N}\}\ .$$

This induces a set of subset $\bar{\mathsf{U}}_n \subset \mathsf{U}_n$ defined such that a subset $U \in \bar{\mathsf{U}}_n$ is identified with its starting index, i.e for all $i \leq |\bar{\mathsf{U}}_n|$, $\bar{\mathsf{U}}_n \ni U_i := \{i, i+1, \ldots, i+n-1\}$. Indeed,

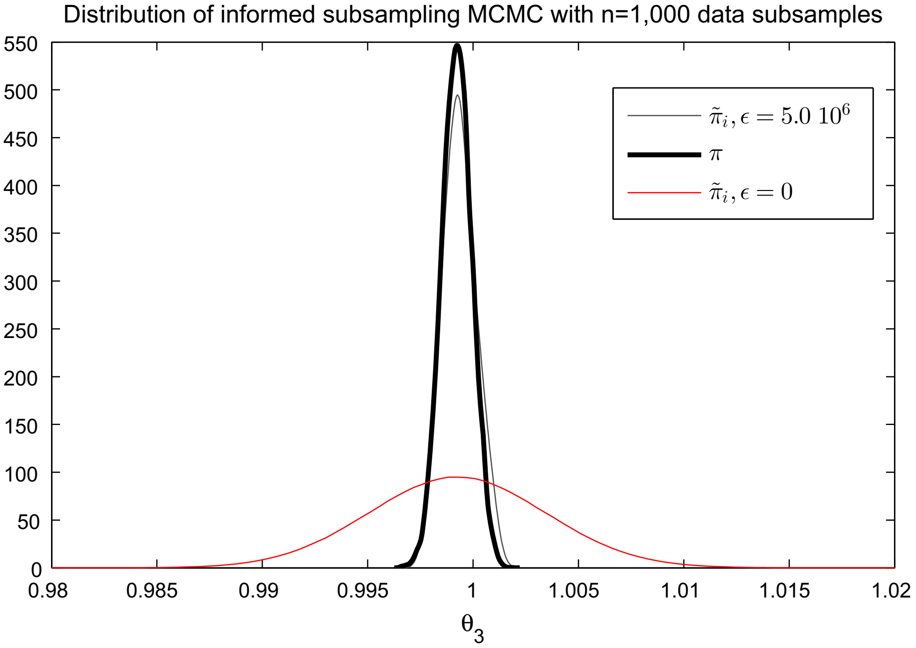


Fig. 5 (Example 3: Autoregressive time series) Inference of the noise parameter with ISS-MCMC, using subsets comprising of $n = 1000$ contiguous time steps of a $N = 10^6$ time-series. The plot represents the distributions $\tilde{\pi}_i$ ($i = 50{,}000$) of the Informed Sub-Sampling Markov chain for two different values of $\epsilon \in \{0, 5.0\ 10^6\}$. These distributions where obtained from the replication of 1000 independent chains

using such subsamples yields a tractable likelihood (37) as otherwise, missing variables need to be integrated out, hence loosing the simplicity of our approach.

With some abuse of notation, the proposal kernel R can be written as a transition kernel on the alphabet $\{0, \ldots, N - n + 1\}$. It is defined in this example as:

$$R(i;\, j) = \mathbb{1}_{i \neq j} \left\{ \omega \frac{\exp(-\lambda|j-i|)}{\sum_{j \leq |\bar{\mathsf{U}}_n|,\, j \neq i} \exp(-\lambda|j-i|)} + (1-\omega)\frac{1}{|\bar{\mathsf{U}}_n| - 1} \right\} . \qquad (39)$$

The rationale is to propose a new subset through a mixture of two distributions: the first gives higher weight to local moves and the latter allows jumps to remote sections of the time series. In this example, we have used $\omega = 0.9$ and $\lambda = 0.1$. We study the efficiency of ISS-MCMC in function of n, ϵ and S.

For any subsample Y_U, $U \in \bar{\mathsf{U}}_n$, we have set the summary statistics $\bar{S}(Y_U)$ to the solution of the AR(2) Yule-Walker equations for the dataset Y_U. As shown in Fig. 4, this choice of summary statistics satisfies (graphically) Assumption **A.**4 with $\gamma \approx 5\ 10^6$. We therefore set $\epsilon = 5.0\ 10^6$ to make sure that A_n (27) is bounded. Theoretically, Proposition 5 guarantee that the bias is controlled. This is illustrated graphically in Fig. 5 where π is compared to $\tilde{\pi}_i$ ($i = 50{,}000$). We also report the distribution of the Informed Sub-Sampling chain when the $\epsilon = 0$, i.e when the subsampling is actually uninformed and all subsamples have the same weight. In the latter case, A_n is not bounded which explains why the bias on $\lim_{i\to\infty} \|\pi - \tilde{\pi}_i\|$ is not controlled. Figures 6 and 7 illustrate the distribution of $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ for some runs of ISS-MCMC with $\epsilon = 0$ and $\epsilon = 5.0\,10^6$. Finally, Fig. 8 gives a hint at the computational efficiency of ISS-MCMC. Metropolis–Hastings was compared to ISS-MCMC with $n \in \{1000;\, 5000;\, 10{,}000\}$ and $\epsilon \in \{0;\, 1;\, 5.0\,10^6\}$. The performance indicator is defined as the average of the marginals Total Variation distance, i.e

$$\mathrm{TV}(t) = \frac{1}{d}\sum_{j=1}^{d} \|\pi^{(j)} - \tilde{\pi}_t^{(j)}\| ,$$

where $\pi^{(j)}$ and $\tilde{\pi}_t^{(j)}$ are respectively the true j-th marginal and the j-th marginal of the chain distribution after a runtime of t seconds. The true marginals were estimated from a long Metropolis–Hastings chain, at stationarity. $\tilde{\pi}_t^{(j)}$ was estimated using 500 independent chains starting from the prior. The Matlab function `ksdensity` in default settings was applied to estimate $\|\pi^{(j)} - \tilde{\pi}_t^{(j)}\|_{\mathrm{TV}}$ from the chains samples, hence the variability. On the one hand, when $\epsilon = 0$, there is no informed search for subsamples which makes the algorithm much faster than the other setups but yields a significant bias (larger than 0.5). On the other hand, setting $\epsilon > 0$ adds to the computational burden but allows to reduce the bias. In fact, for $n = 5000$ and a computational budget of $t = 1000\,$s, the bias of ISS-MCMC is similar to that of Metropolis Hastings but converges 100 times as fast. Finally, note that as expected by the theory, when the unrepresentative subsamples are not penalized enough (e.g. by setting $\epsilon = 1$), ISS-MCMC yields a significant bias which hardly improves on uninformed subsampling when $n = 10{,}000$. Following Assumption **A.**4, we see that setting $\epsilon = 5.0\,10^6$

Fig. 6 (Example 3: Autoregressive time series) Marginal distribution of θ_3 and distribution of $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ for four independent Informed Sub-Sampling Markov chains with $\epsilon = 0$ and $n = 1000$

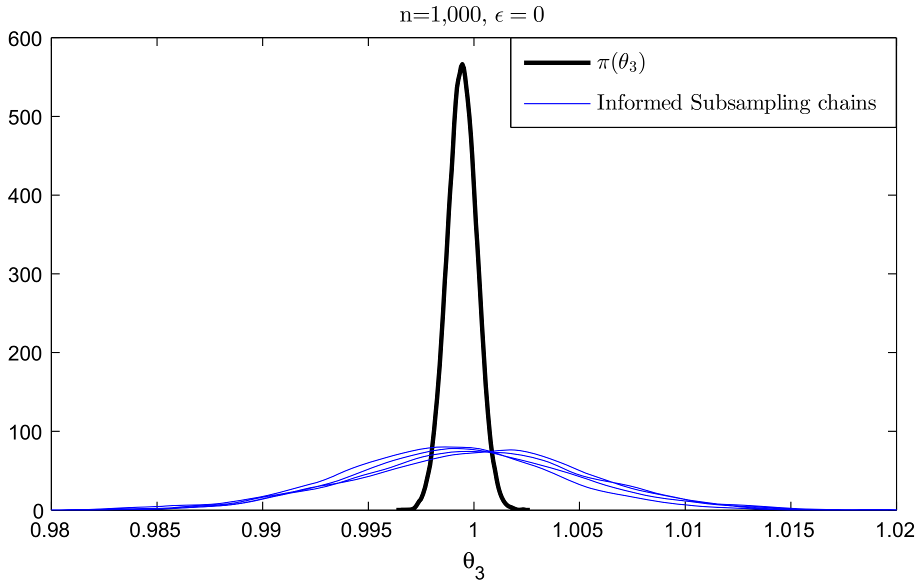
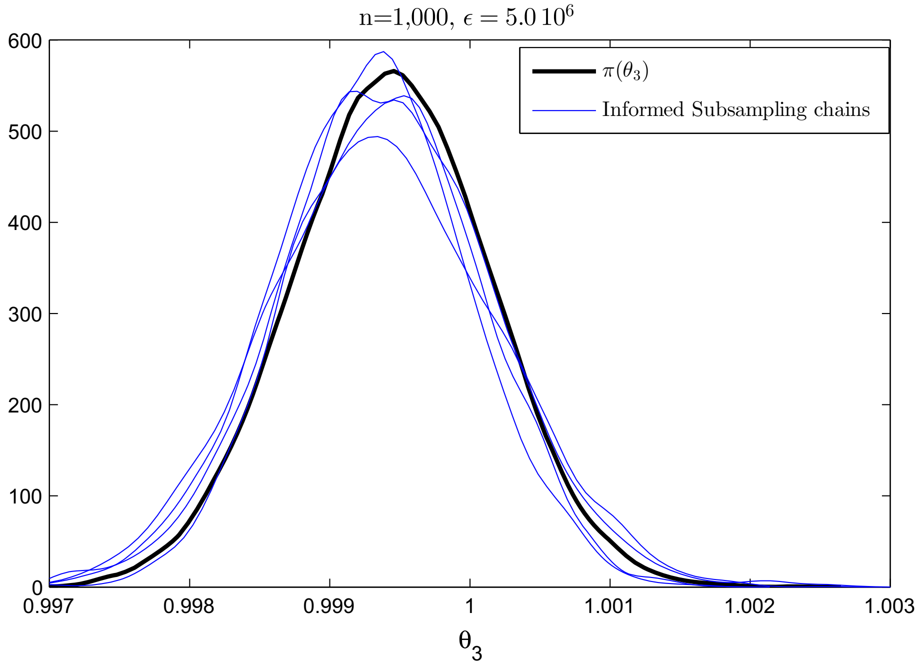


Fig. 7 (Example 3: Autoregressive time series) Marginal distribution of θ_3 and distribution of $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ for four independent Informed Sub-Sampling Markov chains with $\epsilon = 5.0\,10^6$ and $n = 1000$



significantly reduces the bias. Note that setting $\epsilon > 5.0\,10^6$ could potentially reduce further the bias but may fail the algorithm: indeed, when ϵ is too large the chain $\{U_k,\ k \in \mathbb{N}\}$ gets easily stuck on a set of the best subsamples (for this choice of summary statistics) and may considerably slow down the convergence of the algorithm.

Other choices of summary statistics can be considered. Since Y is modelled as an autoregressive time series, an option would be to set the summary statistics as the empirical autocorrelation function. Fig. 9 shows that it is not a recommended choice. The left panel suggests that Assumption **A.**4 does not hold for this type of summary statistics: good subsets yield a large value for $\log f(Y \mid \theta) - (N/n) \log f(Y_U \mid \theta)$ and conversely for bad subsets, hence generating a bias (see Eq. (30)). As a consequence, the right panel which shows a clear mismatch between $\pi^{(3)}$ and the Informed Sub-Sampling third marginal is not surprising.

6.3 Logistic regression example

Example 4 A d-dimensional logistic regression model is parameterized by a vector $\theta = (\theta_1, \ldots, \theta_d) \in \Theta \subset \mathbb{R}^d$. Observations are realizations of the following model:

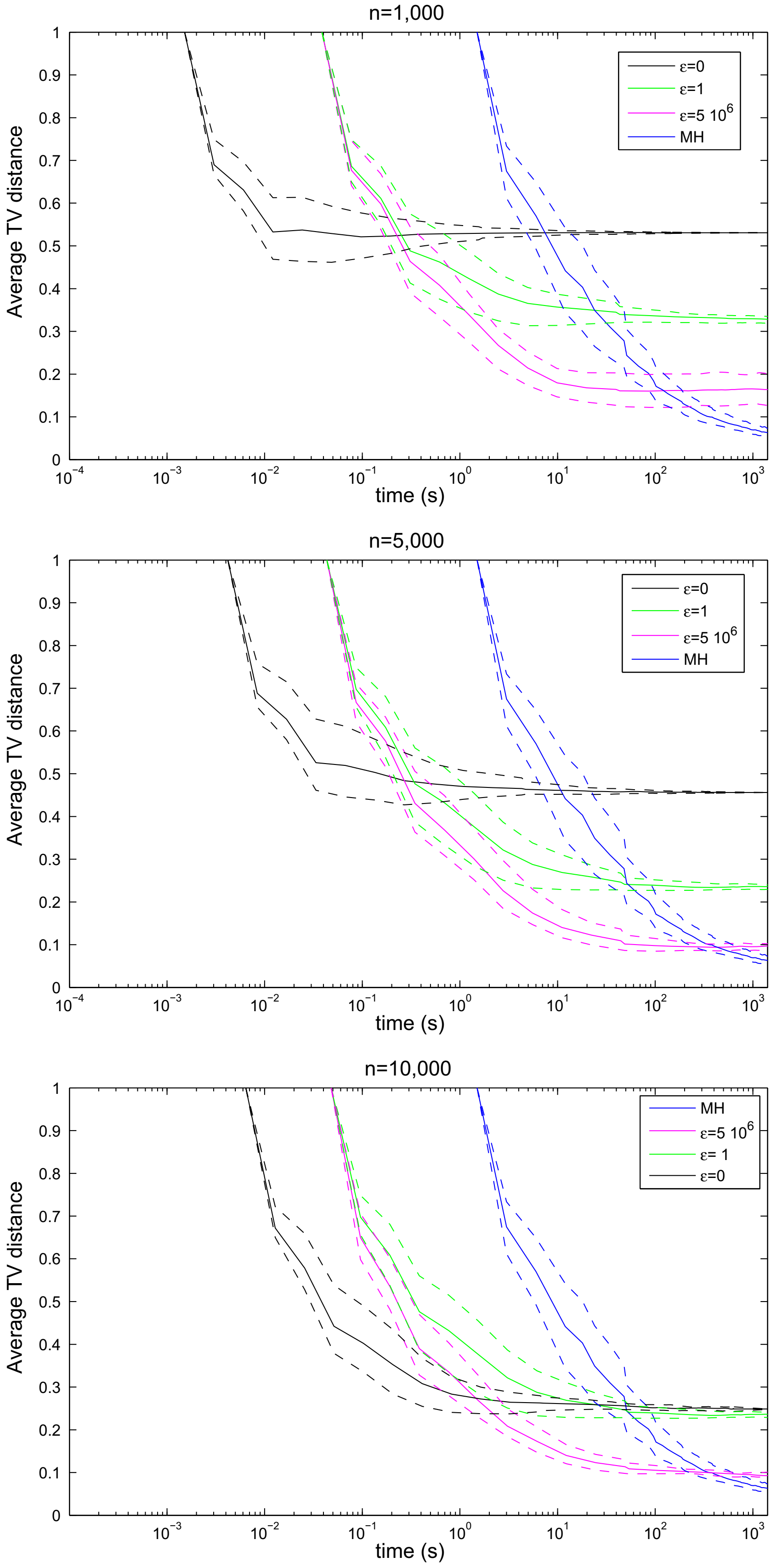


Fig. 8 (Example 3: Autoregressive time series) Average Total variation distance over the three marginals between π and $\tilde{\pi}_t$ in function of the simulation time t. The dashed lines represent the first and third quartiles. Scenario $n = 1000$ (top), $n = 5000$ (middle) and $n = 10{,}000$ (bottom) with three different ϵ. Note that in all the three plots, the M–H curves are identical and are just reported for comparison purpose

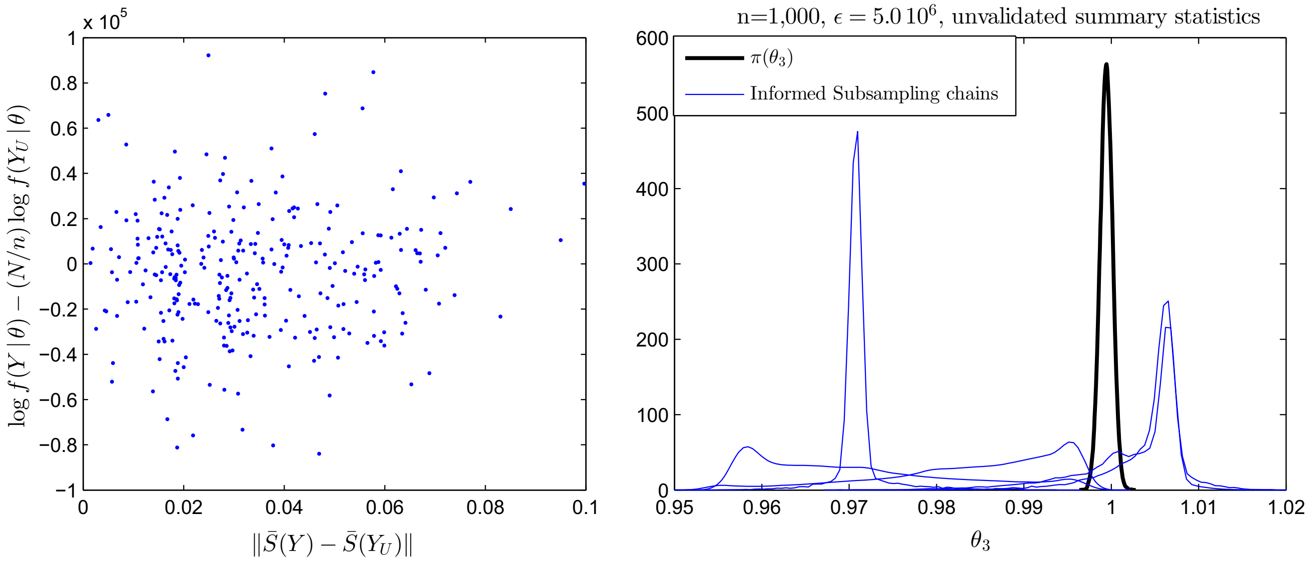


Fig. 9 (Example 3: Autoregressive time series) In this case, the summary statistics were defined as the first 5 empirical autocorrelation coefficients. The left panel shows that this is not a recommended choice and the right panel illustrates the distribution of $\{\tilde{\theta}_i,\ i \in \mathbb{N}\}$ for four independent Informed Sub-Sampling Markov chains ($\epsilon = 5.0\,10^6$, $n = 1000$ and this choice of summary statistics), yielding an obvious mismatch

- simulate covariates $X_i = (X_{i,1}, \ldots, X_{i,d}) \sim \mathcal{N}(0, (1/d)^2)$
- simulate Y_i given θ and X_i as

$$Y_i = \begin{cases} 1 & \text{w.p. } 1/\left(1 + e^{-\theta X^T}\right) , \\ 0 & \text{otherwise} . \end{cases} \tag{40}$$

We have simulated $N = 10^6$ observations $Y_1, Y_2, \ldots$ under the true parameter $\theta^* = (1, 2, -1)$ $(d = 3)$.

The summary statistics were set as the maximum likelihood estimator returned by the Matlab routine `glmfit` and were graphically validated, as in Fig. 4. The tolerance parameter was consequently set $\epsilon = 5.0\,10^6$. We study the influence of n on the Informed Sub-Sampling chain marginal distributions in Fig. 10. We note that as soon as $n \geq 5000$, the bias vanishes and that when random subsampling is used (i.e $\epsilon = 0$), the bias is much larger. Of course, Fig. 10 only gives information about the marginal distributions. To complement the study, we consider estimating the probability $\pi(D)$ where D is the domain defined as:

$$D \subset \Theta = \{\theta_1 \in (0.98, 1.00)\,;\ \theta_2 \in (1.98, 2.01)\\ \theta_3 \in (-0.98, -0.95)\}\ ,$$

in order to check that the joint distribution π is reasonably inferred. Numerical integration using a long Metropolis–Hastings algorithm, gave the ground truth $\pi(D) = 0.1$. The top panel of Fig. 11 illustrates the Monte Carlo estimation of $\pi(D)$ based on $i = 10{,}000$ iterations of ISS-MCMC implemented with $n \in \{1000\,;\ 5000\,;\ 10{,}000\}$ and compares it to Metropolis–Hastings. As expected ISS-MCMC has a negligible bias and the variance of the estimator decreases when n increases. Indeed, when n increases, the Informed Sub-Sampling process is less likely to pick irrelevant subsets, which in turns lower the variability of the chain. The Monte-Carlo estimation based on ISS-MCMC with $n = 10{,}000$ and Metropolis–Hastings are very similar. However, when we normalize the experiment by the CPU time, Metropolis–Hastings is clearly outperformed by ISS-MCMC. The lower panel of Fig. 11 assumes that only $t = 500$ s of computation are available. All the chains are started from θ^*. Table 3 reports the quantitative details of this experiment. In such a situation, one should clearly opt for the Informed Sub-Sampling approach as the Metropolis–Hastings algorithm only achieves 50 iterations for this amount of computation and as such fails to reach stationarity.

We also compare ISS-MCMC with two other algorithms that approximate the Metropolis-Hastings algorithm by using subset of data, drawn, unlike ISS-MCMC, uniformly at random. More precisely, we have implemented the Stochastic Gradient Langevin Dynamic (SGLD) from Welling and Teh (2011) and the Subsampled likelihoods M–H (SubLhd1) algorithm from Bardenet et al. (2014) along with an improved version of this algorithm that makes use of control variates, referred to as the Improved Confidence sampler in Bardenet et al. (2017) but abbreviated here as SubLhd2 for simplicity. Those algorithms have been implemented in their default

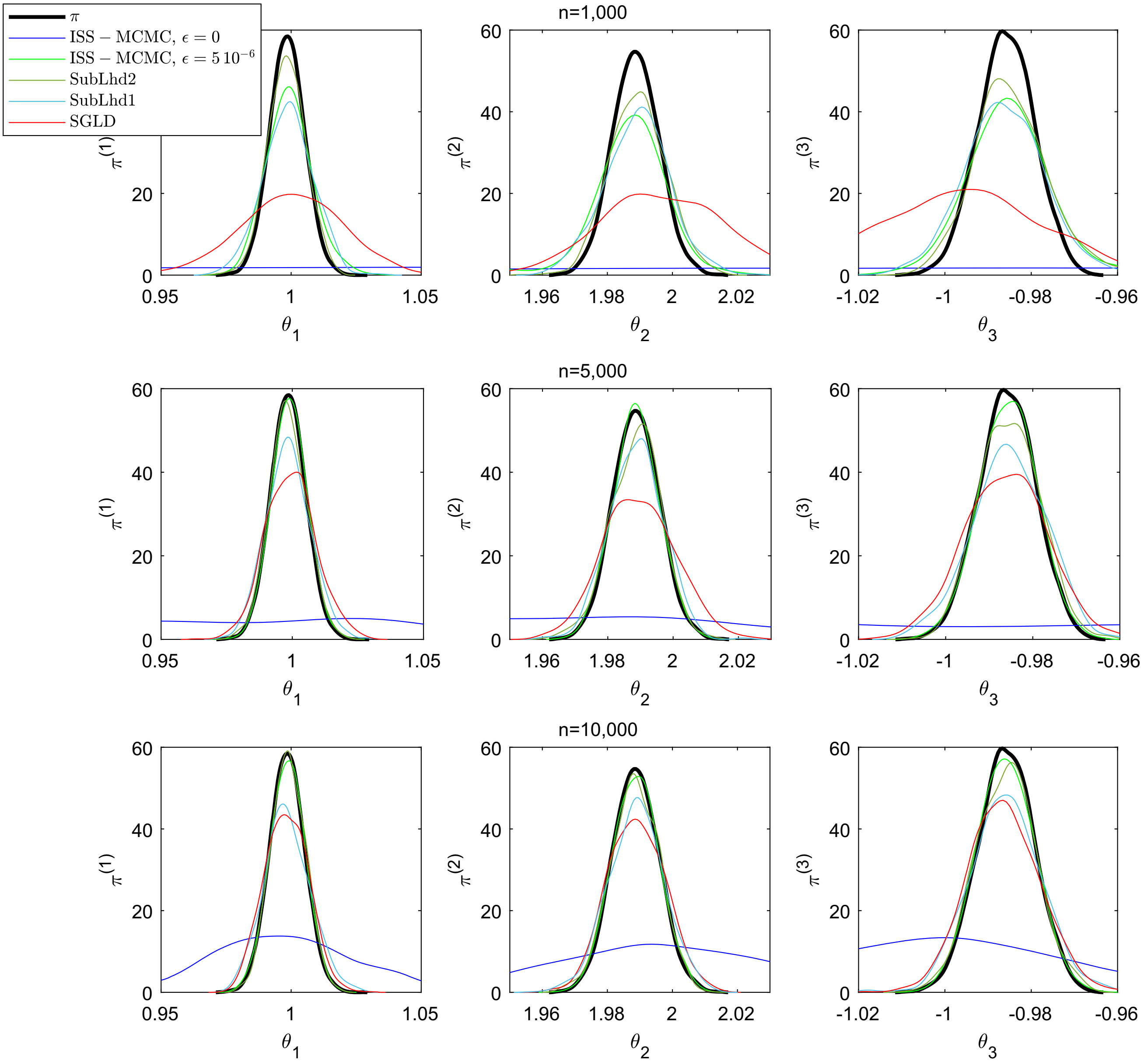
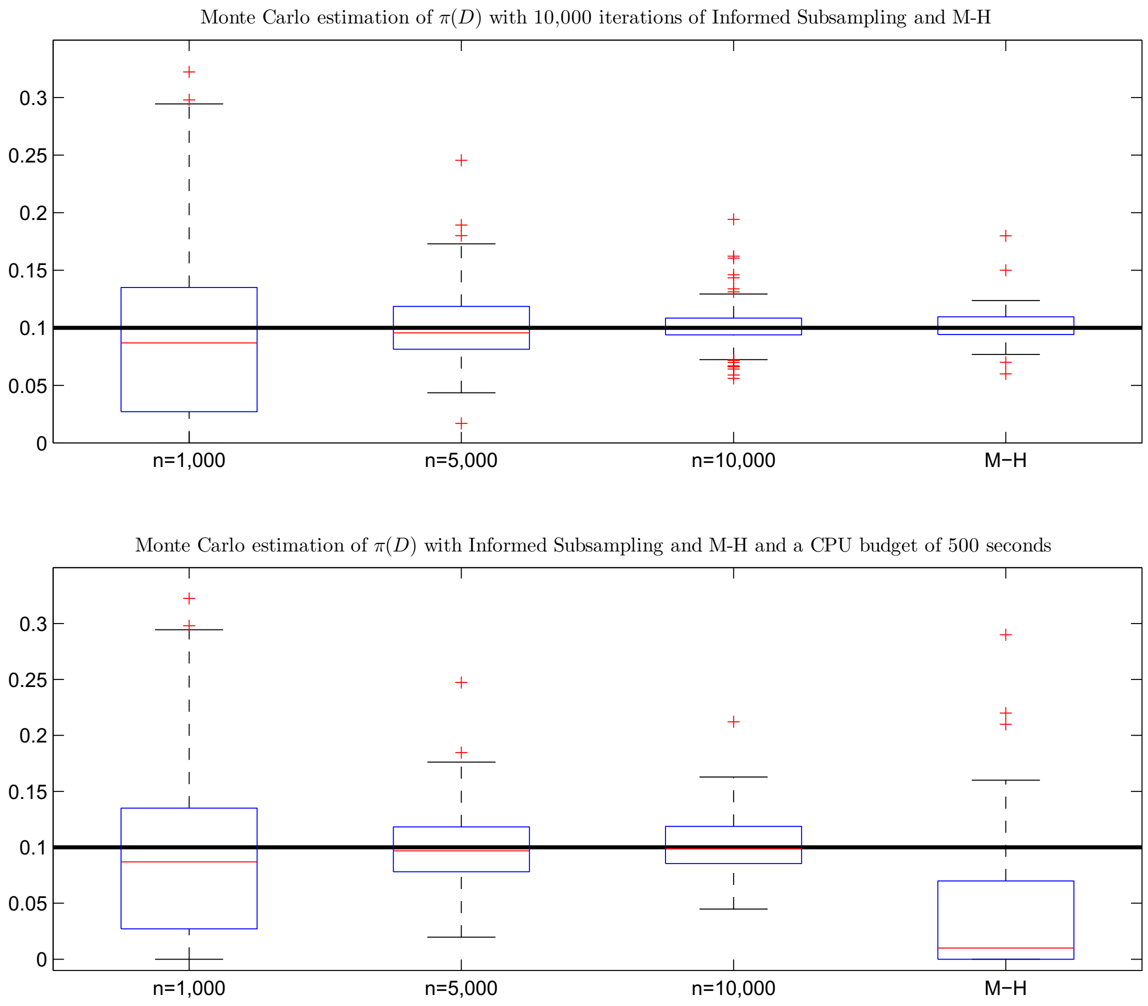


Fig. 10 (Example 4: Logistic regression) Stationary marginal distributions of several algorithms approximating Metropolis–Hastings using subsamples: ISS-MCMC with $\epsilon = \{0\,;\, 5.0\,10^{4}\}$, the Subsampled likelihood M–H algorithm Bardenet et al. (2014) (SubLhd1) (and its improved version denoted SubLhd2, see Bardenet et al. (2017)) and the Stochastic Gradient Langevin Dynamic (SGLD) Welling and Teh (2011). The plots represent the marginal distribution $\tilde{\pi}_i$, (after $i = 1000$ iterations) and different subset sizes $n \in \{1000\,;\, 5000\,;\, 10{,}000\}$. The true marginal π is in black. $\tilde{\pi}_i$ was estimated by simulating 1000 iid copies of the Markov chain generated by the five algorithms

version, following the parameterization prescribed in their original article. All those methods are inexact and we are interested in comparing the bias/variance tradeoff per CPU time unit. Results in terms of convergence in distribution and Monte Carlo estimation are reported respectively in Fig. 10 and Table 3. For this model, SGLD and SubLhd1 show a larger bias than ISS-MCMC and SubLhd2: they need larger subset size n to achieve a similar precision than ISS-MCMC or SubLhd2, see Fig. 10. SubLhd2 seems to outperform ISS-MCMC when n is low in terms of distribution bias but the two methods perform equally good when $n \geq 5000$. Quantitatively, the Monte Carlo estimation of $\pi(D)$ appears better with SubLhd2 than any other method for any subset size, as indicated by the RMSE reported at Table 3. However, looking at the comparative boxplot representing the distribution of the Monte Carlo estimator of $\pi(D)$ in time normalized experiments (Fig. 12), one can see that when n is larger than 5000, estimators from ISS-MCMC and SubLhd2 are quite


Monte Carlo estimation of $\pi(D)$ with 10,000 iterations of Informed Subsampling and M-H

Monte Carlo estimation of $\pi(D)$ with Informed Subsampling and M-H and a CPU budget of 500 seconds

Fig. 11 (Example 4: Logistic regression) Estimation of $\pi(D)$ based on ISS-MCMC implemented with $n \in \{1000\,;\,5000\,;\,10{,}000\}$ and Metropolis–Hastings. Top: the experiment is iteration-normalized, i.e the chains run for 10,000 iterations. Bottom: the experiment is time-normalized, i.e the chains run for 500 s. Each chain was replicated 100 times and started from θ^*

similar confirming the qualitative impression of Fig. 10 and perhaps moderating the RMSE-based assessment made at Table 3.

6.4 Binary classification

Example 5 A training dataset consisting of $N = 10^7$ labeled observations $Y = \{Y_k,\ k \le N\}$ from a 2 dimensional Gaussian mixture model is simulated with

$$Y_k \mid I_k = i \sim \mathcal{N}(\mu_i, \Gamma_i)\,, \qquad I_k \sim \text{Bernoulli}(1/2)\,,$$

where $\mu_1 = [\theta_1\,,\,0]$, $\mu_2 = [\theta_2\,,\,0]$, $\Gamma_1 = \text{diag}([\theta_3/2\,,\,\theta_3])$ and $\Gamma_2 = \text{diag}([\theta_4/2\,,\,\theta_4])$. We define $\theta = (\theta_1, \theta_2)$ with $\theta_i \in \mathbb{R} \times \mathbb{R}_*^+$ for each model $i \in \{1, 2\}$. A prior distribution $(\theta_1, \theta_2) \sim_{i.i.d.} p := \mathcal{N}(0, 1/2) \otimes \Gamma(1, 2)$ ($\Gamma(a, b)$ is the Gamma distribution with shape a and rate b) is assigned to θ. Consider an algorithm a that simulates a Markov chain

$$\theta_{\mathrm{a}} := \left\{\left(\theta^{(\mathrm{a})}_{1,k}, \theta^{(\mathrm{a})}_{2,k}\right), \quad k \in \mathbb{N}\right\}$$

targeting the posterior distribution of θ given Y, perhaps approximately. We consider the real-time supervised classifier $I^*_{\mathrm{a}}(t)$, driven by θ_{a}, for the test dataset $Y^* = \{Y^*_k,\ k \le N_{\text{test}}\}$ ($N_{\text{test}} = 10^4$) and defined as:

$$\begin{aligned} &I^*_{\mathrm{a}}(t) = (I^*_{\mathrm{a},1}(t), \ldots, I^*_{\mathrm{a},N_{\text{test}}}(t))\,, \\ &I^*_{\mathrm{a},k}(t) = \arg\max_{i\in\{1,2\}} f(Y^*_k \mid \bar{\theta}^{(\mathrm{a})}_{i,\kappa_{\mathrm{a}}(t)})\,, \end{aligned} \qquad (41)$$

where $\kappa_{\mathrm{a}}(t) = \sup_{k\in\mathbb{N}}\{\tau^{\mathrm{a}}_k \le t\}$ and $\bar{\theta}^{(\mathrm{a})}_{i,k} = (1/k)\sum_{\ell=1}^{k} \theta^{(\mathrm{a})}_{i,\ell}$. We have defined τ^{a}_k as the wall clock time to generate k iterations of algorithm a. We define the live classification error rate as $\epsilon_{\mathrm{a}}(t) = \|I^*_{\mathrm{a}}(t) - I^*\|_1$ where $I^*_{\mathrm{a}} = (I^*_{\mathrm{a},1}, \ldots, I^*_{\mathrm{a},N})$

Table 3 (Example 4: Logistic regression) Tradeoff Bias-Variance of the Monte Carlo estimator from Metropolis–Hastings, ISS-MCMC, Stochastic Gradient Langevin Dynamics (SGLD) Welling and Teh (2011), the Subsampled likelihoods (SubLhd1) Bardenet et al. (2014) and the improved Confidence Sampler (SubLhd2) Bardenet et al. (2017) for a fixed computational budget of 500 s

Algorithm	Time/iter.(s)	Iter. completed	RMSE	$\text{var}\{\widehat{\pi(D)}\}$
M–H	10	50	0.1417	0.004
ISS-MCMC, $n = 1000$	0.05	10,000	0.1016	0.0104
ISS-MCMC, $n = 5000$	0.08	6250	0.0351	0.0012
ISS-MCMC, $n = 10{,}000$	0.13	3840	0.0267	0.0007
SGLD, $n = 1000$	0.08	6000	0.1370	0.0157
SGLD, $n = 5000$	0.11	5250	0.0996	0.0100
SGLD, $n = 10{,}000$	0.12	4500	0.0326	0.0011
SubLhd1, $n = 1000$	1.45	350	0.0762	0.0042
SubLhd1, $n = 5000$	1.56	323	0.0680	0.0046
SubLhd1, $n = 10{,}000$	2.24	223	0.0656	0.0044
SubLhd2, $n = 1000$	0.10	5046	0.0304	0.0002
SubLhd2, $n = 5000$	0.14	3581	0.0260	0.0006
SubLhd2, $n = 10{,}000$	0.19	2631	0.0195	0.0002

Those results were replicated using 100 replications of each algorithm. Note that for SubLhd1 and SubLhd2, n corresponds to the initial subset size and not to the actual subsample size that was actually used in each iteration, a parameter which is chosen by the algorithms

and I_k^* is the true class of Y_k^*. We compare $\epsilon_{\mathtt{a}}$ for three different algorithms $\mathtt{a}$: ISS-MCMC, Metropolis–Hastings and Subsampled Likelihoods (SubLhd1) (Bardenet et al. 2014).

In this simulation example, we have used the true value $\theta^* = (-1, 1/2, 1, 1/2)$ and simulated Y such that it contains the same number of observations from model 1 and model 2 i.e $N/2$. The three algorithms were implemented with the same proposal kernel, namely a single site random walk with adaptive variance that guarantees an acceptance rate between 0.40 and 0.50, see Roberts et al. (2001); Haario et al. (2001). ISS-MCMC was implemented with parameters $n = 1000$ and $\epsilon = 10^7$. The summary statistics were taken as $S(Y_U) = 0$ if $\sum_{k\in U} \mathbb{1}_{\{I_k=1\}} \neq \sum_{k\in U} \mathbb{1}_{\{I_k=2\}}$ and

$$\bar{S}(Y_U) = \Bigg[(2/n)\sum_{k=1}^{n/2} Y_k \mathbb{1}_{\{I_k=1\}},\ \text{tr}(\text{cov}(Y_k,\ k\in U,\ I_k=1))\,, \\ (2/n)\sum_{k=1}^{n/2} Y_k \mathbb{1}_{\{I_k=2\}},\ \text{tr}(\text{cov}(Y_k,\ k\in U,\ I_k=2))\Bigg] \tag{42}$$

otherwise. This choice allows to keep the right proportion of data from the two models in any subsample used for the inference. The statistics in (42) are sufficient for each model, taken separately. Subsampled Likelihoods was implemented with the default parameters prescribed in the introduction of Section 4 in Bardenet et al. (2014).

Figure 13 compares the live classification error rate achieved by the three algorithms. We also report the optimal Bayes classifier which achieves $\epsilon_B(t) = 0.0812$ classifying Y_k^* in class 1 if $Y_{k,1}^* < 0$ and in class 2 alternatively. Unsurprisingly, Metropolis–Hastings is penalized because it evaluates the norm of a $N = 10^7$ dimensional vector at each iteration. Subsampled Likelihoods does slightly better than M–H but suffers from the fact that close to stationary regime, the algorithm ends up drawing the quasi-entire dataset with high probability, a fact which was explained in Bardenet et al. (2014).

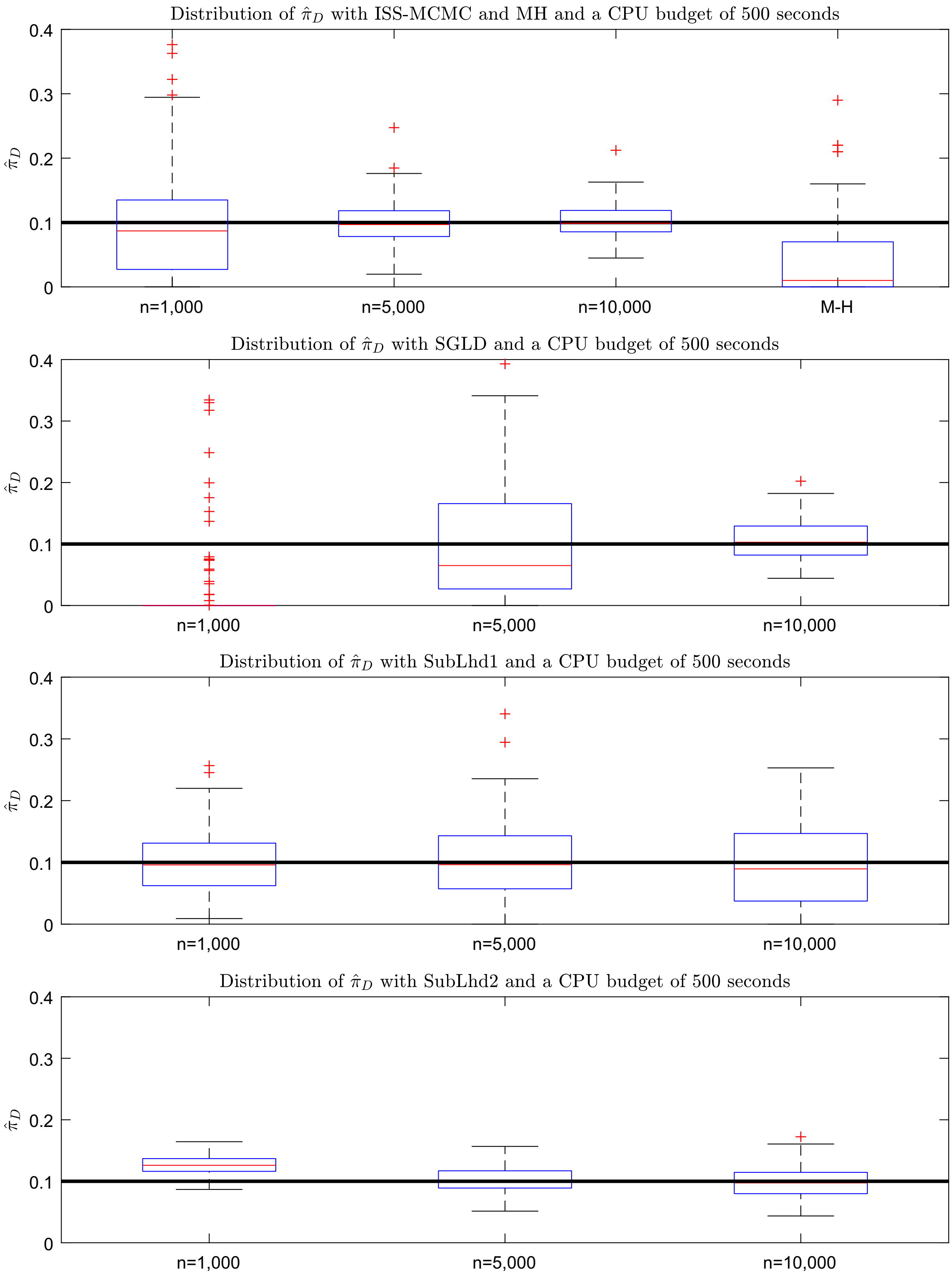
6.5 Additional details for the handwritten digit inference (Example 1)

In the handwritten digit example (Example 1), we have used batches of $n = 100$ data. Since the initial dataset comprises 2000 observations per digit, the summary statistics were defined in a way that any subsample contains 20 observations from each class. More precisely, we have set for any subsample Y_U, $S(Y_U) = 0$ if for at least one class $i \in \{1,\ldots,5\}$, $(1/n)\sum_{k\in U} \mathbb{1}_{\{J(k)=i\}} \neq 20$ and

$$\bar{S}(Y_U) = \left\{\sum_{k\in U} \phi(\theta_{J(k)})\mathbb{1}_{\{J(k)=i\}} \Big/ \sum_{k\in U} \mathbb{1}_{\{J(k)=i\}}\right\}_{i=1}^{5}$$

otherwise. We set the bandwidth to $\epsilon = 10^5$. The proposal kernels of the Informed Sub-Sampling chain and M–H were defined as the same Random Walk kernel. In particular, at each iteration only a bloc of the template parameter of one of the five classes is updated. The variance parameter of the Random Walk is adapted according to the past trajectory of the chain, so as to maintain an acceptance rate of .25.

Figure 14 reports the empirical marginal distribution of one component for each vector $\theta_1,\ldots,\theta_5$ obtained from ISS-MCMC and from M–H. Those distributions are estimated from 50,000 iterations of both algorithms, in stationary


Distribution of $\hat{\pi}_D$ with ISS-MCMC and MH and a CPU budget of 500 seconds

Distribution of $\hat{\pi}_D$ with SGLD and a CPU budget of 500 seconds

Distribution of $\hat{\pi}_D$ with SubLhd1 and a CPU budget of 500 seconds

Distribution of $\hat{\pi}_D$ with SubLhd2 and a CPU budget of 500 seconds

Fig. 12 (Example 4: Logistic regression) Estimation of $\pi(D)$ based on ISS-MCMC, MH, SGLD, SubLhd1 and SubLhd2 implemented with $n \in \{1000\,;\ 5000\,;\ 10{,}000\}$. The Monte Carlo estimation of $\pi(D)$ was carried out using those algorithms for 500 s. Each estimation was replicated 100 times and started from θ^* for each algorithm. Note that for SubLhd1 and SubLhd2, n corresponds to the initial subset size and not to the actual subsample size that was actually used in each iteration, a parameter which is adaptively tuned by the algorithms

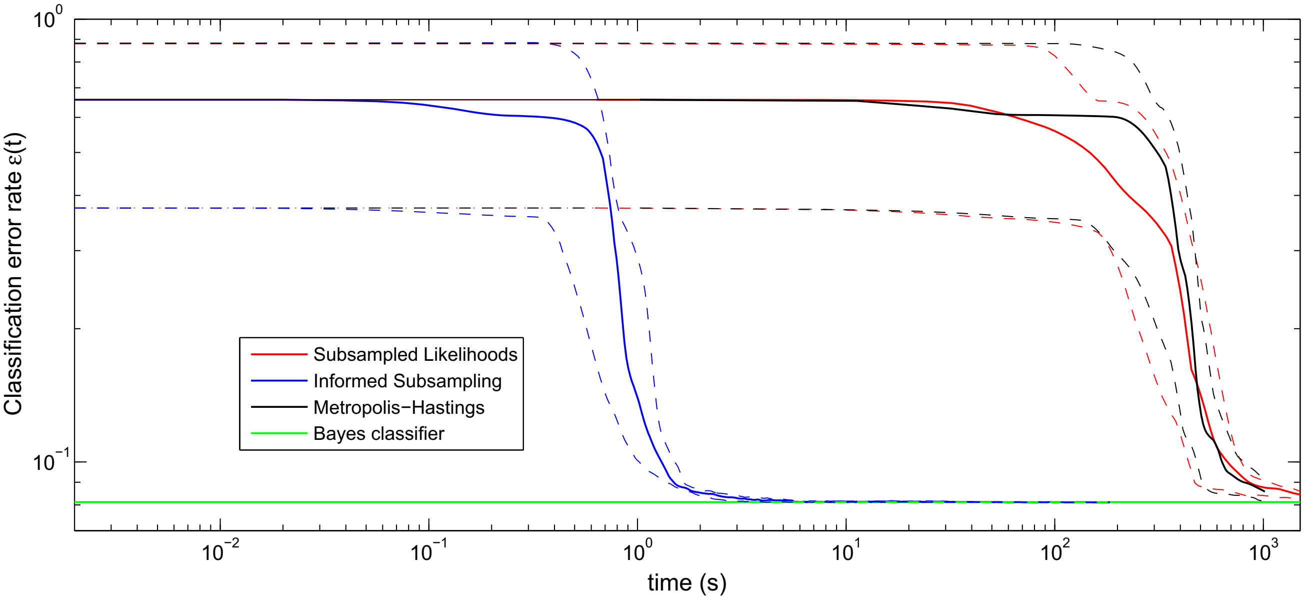


Fig. 13 (Example 5: Binary classification) Live classification error rate for three algorithms. This plot was generated by classifying the same test dataset Y^* using the same training dataset Y for the three algorithms. The variability arises from the initial state of the Markov chains. We have used 30 different initial states for the three algorithms and report the median (plain line) and the two quartiles (dashed lines)

regime. This shows that the distribution of those parameters are in line with each other.

7 Conclusion

When the available computational budget is limited, inferring a statistical model based on a tall dataset in the Bayesian paradigm using the Metropolis–Hastings (M–H) algorithm is not computationally efficient. Several variants of the M–H algorithm have been proposed to address this computational issue (Bardenet et al. 2014; Banterle et al. 2015; Korattikara et al. 2014; Maclaurin and Adams 2015). However, (i) they often lose the original simplicity of M–H, (ii) they are only applicable in situations where the data are independent and (iii) the computational cost of one iteration is stochastic which can potentially compromise any computational saving. The method presented in this paper, Informed Sub-Sampling MCMC, pushes the approximation one step forward: the computational cost of one iteration is deterministic and is controlled through a user specified parameter, the size of the subsamples.

The aforementioned methods rely on subsampling the whole dataset uniformly at random, at each iteration. Using such subsamples in lieu of the whole dataset in the original M–H chain leads to an algorithm whose approximation of π comes with no guarantee when the subsamples size is fixed. Our main contribution is to show that assigning a distribution to the subsamples that reflects their fidelity to the whole dataset allows one to control the L1 approximation error, even when the subsample size is fixed, which is of computational interest.

Informed Sub-Sampling MCMC, offers an alternative to situations where other scalable Metropolis–Hastings variants cannot be implemented (because the model does not satisfy the assumptions e.g. independence of the data, existence of a concentration inequality for the model or a cheap lower bound of the likelihood) or are inefficient (because the method ends up using nearly the whole dataset at each iteration). However, in scaling up Metropolis–Hastings there is no free lunch neither. In particular, our method replaces the uniform subsampling approach by a more sophisticated subsampling mechanism involving summary statistics. In this regard, even though our method is in principle widely applicable, it will only be useful in situations where a cheap summary statistics function satisfying Assumption **A.**4 is available. In particular, we have shown that our method will give meaningful results when the maximum likelihood estimator is cheap to compute, which somewhat correlates with the optimality of summary statistics in ABC established in Fearnhead and Prangle (2012). We note that it is possible to construct a counterpart to ISS-MCMC in the Sequential Monte Carlo framework, as a sequence of scaled subposteriors naturally arises in our algorithm. We leave the design and evaluation of such a method for future research.

Similarly to the other noisy subsampling methods that approximate MH (e.g. Welling and Teh (2011), Korattikara et al. (2014), Bardenet et al. (2017), etc.), our theoreti-

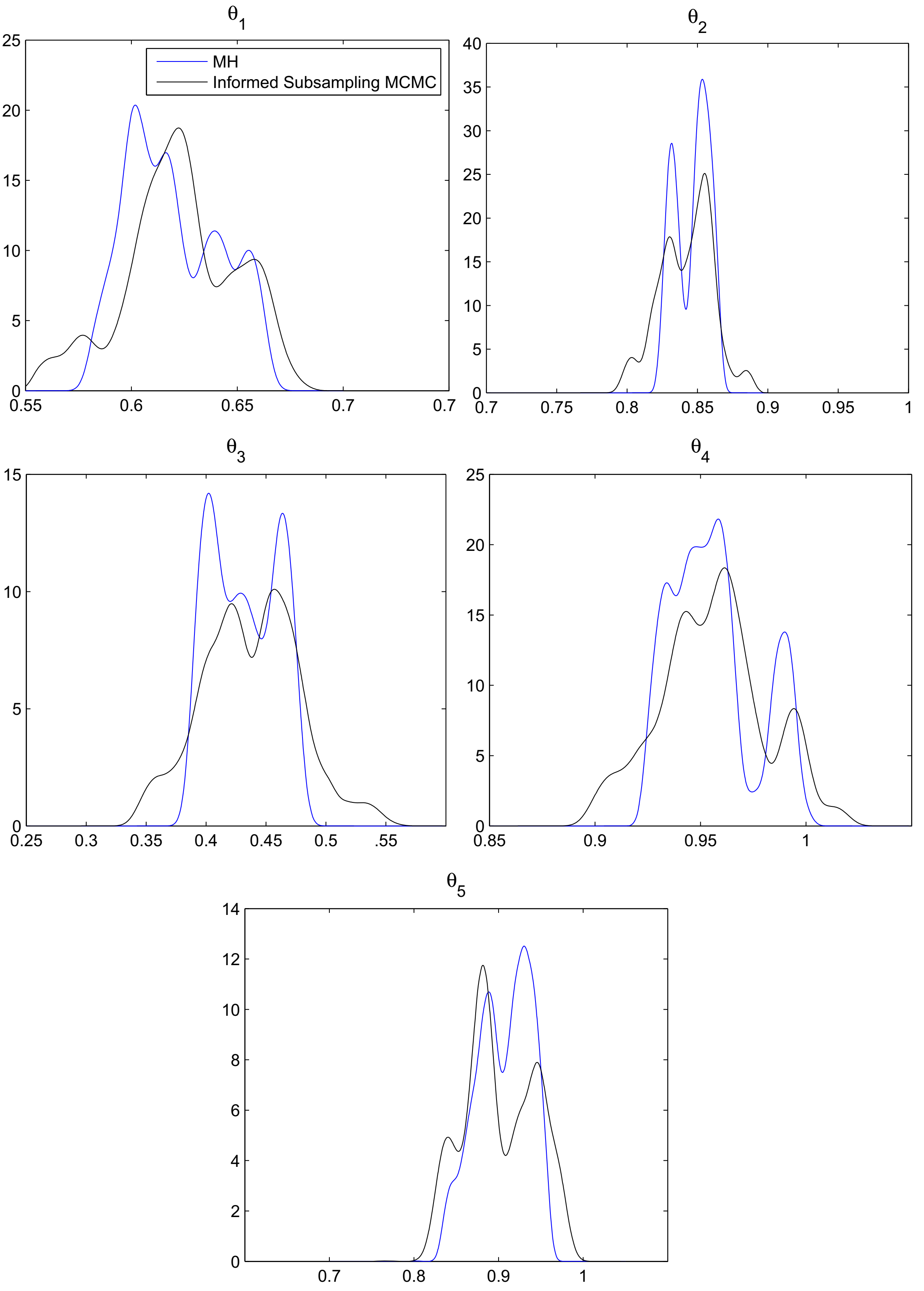


Fig. 14 (Example 1: Handwritten digits) Empirical marginal distribution of one component of the vectors $\theta_1, \ldots, \theta_5$ using the Metropolis–Hastings chain (blue) and the Informed Sub-Sampling chain (black), estimated from 10,000 transitions at stationary regime

cal results address the convergence in distribution of the ISS-MCMC chain to the posterior distribution. Questions of interest to practitioners encompass establishing Law of Large Numbers and Central limit theorems for those types of algorithms. Addressing those questions would allow a better understanding of approximate MCMC applied to large dataset contexts, from a practical perspective. In the noisy Monte Carlo literature, the closest result is perhaps Theorem 2.5 of Johndrow et al. (2015) which comes in the form of a non-asymptotic error bound in the L2 norm between πf and its MCMC estimate. The assumptions are however quite strong which prevent applying it to ISS-MCMC and we will study those questions in a future work. Finally, recent developments in the understanding of approximate Markov chains have been carried out in Rudolf and Schweizer (2018). These authors provide explicit convergence bounds in the Wasserstein metric under the Wasserstein ergodicity assumption, a notion which is closely related to geometric ergodicity in V-norm, hence milder than uniform ergodicity. We regard this contribution as a very promising research avenue that unveils new ways of deriving quantitative error bounds for the practical problem of approximate Metropolis–Hastings algorithms that make use of subsets of data, under more realistic assumptions.

Acknowledgements The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289. Nial Friel's research was also supported by an Science Foundation Ireland grant: 12/IP/1424. Pierre Alquier's research was funded by Labex ECODEC (ANR - 11-LABEX-0047) and by the research programme New Challenges for New Data from LCL and GENES, hosted by the Fondation du Risque. We thank the Associate Editor and two anonymous Referees for their contribution to this work.

Appendix A Proofs

Appendix A.1 Proof of Proposition 1

Proof For notational simplicity and without loss of generality, we take here g as the identity on Θ. Let $n < N$ and U be a subset of $\{1, \ldots, N\}$ with cardinal n. Consider the power likelihood:

$$\tilde{f}_n(Y_U \mid \theta) = f(Y_U \mid \theta)^{N/n} = \left\{ \prod_{k \in U} f(Y_k \mid \theta) \right\}^{N/n} = \frac{\exp\left\{(N/n) \sum_{k\in U} S(Y_k)\right\}^T \theta}{L(\theta)^N},$$

and the corresponding power posterior:

$$\tilde{\pi}_n(\theta \mid Y_U) = \frac{\exp\left\{(N/n) \sum_{k\in U} S(Y_k)\right\}^T \theta}{L(\theta)^N} p(\theta) \Big/ \tilde{Z}_n(Y_U),$$

where

$$\tilde{Z}_n(Y_U) = \int p(\mathrm{d}\theta) \frac{\exp\left\{(N/n) \sum_{k\in U} S(Y_k)\right\}^T \theta}{L(\theta)^N}.$$

For any θ such that $p(\theta) \neq 0$, write:

$$\log \frac{\pi(\theta \mid Y_{1:N})}{\tilde{\pi}_n(\theta \mid Y_U)} = \left\{ \sum_{k=1}^N S(Y_k) - (N/n) \sum_{k\in U} S(Y_k) \right\}^T \theta + \log \frac{\tilde{Z}_n(Y_U)}{Z(Y_{1:N})}. \tag{A.1}$$

and the KL divergence between $\pi(\,\cdot \mid Y_{1:N})$ and $\tilde{\pi}(\,\cdot \mid Y_U)$, denoted $\mathrm{KL}_n(U)$, simply writes

$$\mathrm{KL}_n(U) = \Delta_n(U)^T \mathbb{E}_\pi(\theta) + \log \frac{\tilde{Z}_n(Y_U)}{Z(Y_{1:N})}, \tag{A.2}$$

where $\Delta_n(U) = \sum_{k=1}^N S(Y_k) - (N/n) \sum_{k\in U} S(Y_k)$. Now, note that

$$\begin{aligned}\tilde{Z}_n(Y_U) &= \int p(\mathrm{d}\theta) \frac{\exp\left\{(N/n) \sum_{k\in U} S(Y_k)\right\}^T \theta}{L(\theta)^N} \\ &= \int p(\mathrm{d}\theta) \frac{\exp\left\{\sum_{k=1}^N S(Y_k) - \Delta_n(U)\right\}^T \theta}{L(\theta)^N} \\ &= \int p(\mathrm{d}\theta) f(Y_{1:N} \mid \theta) \exp\left\{-\Delta_n(U)^T \theta\right\} \\ &= Z(Y_{1:N}) \mathbb{E}_\pi \left\{\exp\left(-\Delta_n(U)^T \theta\right)\right\}.\end{aligned} \tag{A.3}$$

Plugging (A.3) into (A.2) yields:

$$\begin{aligned}\mathrm{KL}_n(U) &= \Delta_n(U)^T \mathbb{E}_\pi(\theta) + \log \mathbb{E}_\pi \left\{\exp\left(-\Delta_n(U)^T \theta\right)\right\}, \\ &= \log \frac{\mathbb{E}_\pi \left\{\exp\left(-\Delta_n(U)^T \theta\right)\right\}}{\exp(-\Delta_n(U)^T \mathbb{E}_\pi(\theta))} \\ &= \log \mathbb{E}_\pi \exp\left[\{\mathbb{E}_\pi(\theta) - \theta\}^T \Delta_n(U)\right].\end{aligned} \tag{A.4}$$

Finally, Cauchy-Schwartz inequality provides the following upper bound for $\mathrm{KL}_n(U)$:

$$\mathrm{KL}_n(U) \le \log \mathbb{E}_\pi \exp\left\{\|\mathbb{E}_\pi(\theta) - \theta\| \, \|\Delta_n(U)\|\right\}. \tag{A.5}$$

□

Appendix A.2 Proof of Proposition 2

Proof Under some weak assumptions, Bernstein-von Mises theorem states that $\pi(\,\cdot \mid Y_{1:N})$ is asymptotically (in N) a Gaussian distribution with the maximum likelihood θ^* as mean and $\Gamma_N = I^{-1}(\theta^*)/N$ as covariance matrix, where

$I(\theta)$ is the Fisher information matrix at θ. Let us denote by Φ the pdf of $\mathcal{N}(\theta^*, \Gamma_N)$. Under this approximation, $\mathbb{E}_\pi(\theta) = \theta^*$ and from (A.3), we write:

$$\begin{aligned}
\exp \mathrm{KL}_n(U) &\approx \int \Phi(\mathrm{d}\theta) \exp\left[\{\theta^* - \theta\}^T \Delta_n(U)\right] \\
&= \int \Phi(\theta^* - \theta) \exp\{\theta^T \Delta_n(U)\}\mathrm{d}\theta \\
&= \int \frac{1}{(2\pi)^{(d/2)}|\Gamma_N|^{(1/2)}} \\
&\quad \exp\left\{-(1/2)\theta^T\Gamma_N^{-1}\theta + \theta^T\Delta_n(U)\right\}\mathrm{d}\theta\,, \\
&= \frac{1}{(2\pi)^{(d/2)}|\Gamma_N|^{(1/2)}} \int \exp\left[-(1/2)\left\{\theta^T\Gamma_N^{-1}\theta\right.\right. \\
&\quad \left.\left. -2\theta^T\Gamma_N^{-1}\Gamma_N\Delta_n(U)\right\}\right]\mathrm{d}\theta\,, \\
&= \exp\{(1/2)\Delta_n(U)^T\Gamma_N\Delta_n(U)\}\,, \qquad \text{(A.6)}
\end{aligned}$$

by integration of a multivariate Gaussian density function. Eventually, (A.6) yields the following approximation:

$$\mathrm{KL}_n(U) \approx \widehat{\mathrm{KL}}_n(U) = (1/2)\Delta_n(U)^T\Gamma_N\Delta_n(U)\,. \qquad \text{(A.7)}$$

□

Appendix A.3 Proof of Proposition 3

Proof Let $\mathsf{U}_n \supset A_n(\theta) := \left\{U \in \mathsf{U}_n,\ g(\theta)^T\Delta_n(U) \leq 0\right\}$ and remark that using Cauchy-Schwartz inequality, we have:

$$\begin{aligned}
&\mathbb{E}\left\{\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\right\} \leq \nu_{n,\epsilon}\left\{A_n(\theta)\right\} \\
&\quad + \sum_{U \in \mathsf{U}_n \backslash A_n(\theta)} \nu_{n,\epsilon}(U)\exp\{\|g(\theta)\|\|\Delta_n(U)\|\}\,.
\end{aligned}$$

Now, define $\bar{\Delta}_n(U) := \bar{S}(Y) - \bar{S}(Y_U)$ where $\bar{S}$ is the normalized summary statistics vector, i.e if $U \in \mathsf{U}_n$, $\bar{S}(Y_U) = S(Y_U)/n$. Clearly, when $N \to \infty$, some terms

$$\exp\{\|g(\theta)\|\|\Delta_n(U)\|\} = \exp\{N\|g(\theta)\|\|\bar{\Delta}_n(U)\|\}$$

will have a large contribution to the sum. More precisely, any mismatch between summary statistics of some subsamples $\{Y_U,\ U \in \mathsf{U}_n\backslash A_n(\theta)\}$ with respect to the full dataset will be amplified by the factor N, whereby exponentially inflating the upper bound. However, assigning the distribution $\nu_{n,\epsilon}$ (12) to the subsamples $\{Y_U,\ U \in \mathsf{U}_n\}$, allows to balance out this effect. Indeed, note that

$$\begin{aligned}
&\mathbb{E}\left\{\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\right\} \leq \nu_{n,\epsilon}\{A_n(\theta)\} \\
&\quad + \sum_{U \in \mathsf{U}_n \backslash A_n(\theta)} \exp\{-\epsilon\|\Delta_n(U)\|^2 \\
&\quad + \|g(\theta)\|\|\Delta_n(U)\|\}/Z(\epsilon)\,,
\end{aligned}$$

where $Z(\epsilon) = \sum_{U \in \mathsf{U}_n} \exp\{-\epsilon\|\Delta_n(U)\|^2\}$ and we have, for a fixed n and when $N \to \infty$, that

$$\nu_{n,\epsilon}(U)\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)} \to_{\|\Delta_n(U)\| \to \infty} 0\,.$$

Since g is bounded, then $\mathbb{E}\left\{f(Y\,|\,\theta)/f(Y_U\,|\,\theta)^{N/n}\right\}$ is bounded too. □

Appendix A.4 Proof of Proposition 4

We preface the proof Proposition 4 with five Lemmas, some of which are inspired from Medina-Aguayo et al. (2016). For notational simplicity, the dependence on (n, ϵ) of any ISS-MCMC related quantities is implicit. For all $(\theta, U) \in \Theta \times \mathsf{U}_n$, we denote by $\phi_U(\theta) = f(y_U\,|\,\theta)^{N/n}/f(y\,|\,\theta)$ and recall that $a(\theta, \theta')$ is the (exact) MH acceptance ratio so that $\alpha(\theta, \theta') = 1 \wedge a(\theta, \theta')$. Unless stated otherwise, $\mathbb{E}$ is the expectation taken under $\nu_{n,\epsilon}$. For simplicity, $\tilde{K}_{n,\epsilon}$ is written as $\tilde{K}_n$.

Lemma 1 *For any $(\theta, \theta') \in \Theta^2$, we have*

$$\tilde{\alpha}(\theta, \theta') \leq \alpha(\theta, \theta')\left\{1 \vee \mathbb{E}\frac{\phi_U(\theta')}{\phi_U(\theta)}\right\}\,.$$

Proof This follows from a slight adaptation of Lemma 3.3 in Medina-Aguayo et al. (2016):

$$\begin{aligned}
\tilde{\alpha}(\theta, \theta') &= \mathbb{E}\left\{1 \wedge \frac{f(Y_U\,|\,\theta')^{N/n}p(\theta')Q(\theta', \theta)}{f(Y_U\,|\,\theta)^{N/n}p(\theta)Q(\theta, \theta')}\,\frac{f(Y\,|\,\theta)f(Y\,|\,\theta')}{f(Y\,|\,\theta)f(Y\,|\,\theta')}\right\} \\
&\quad 1 \wedge \left\{a(\theta, \theta')\mathbb{E}\frac{\phi_U(\theta')}{\phi_U(\theta)}\right\} \\
&\leq 1 \wedge \left[a(\theta, \theta')\left\{\mathbb{E}\frac{\phi_U(\theta')}{\phi_U(\theta)} \vee 1\right\}\right] \\
&\leq \alpha(\theta, \theta')\left\{\mathbb{E}\frac{\phi_U(\theta')}{\phi_U(\theta)} \vee 1\right\}\,,
\end{aligned}$$

where we have used Jensen's inequality and the fact that the inequality $1 \wedge ab \leq (1 \wedge a)b$ holds for $a > 0$ and $b \geq 1$. □

Lemma 2 *For any $\theta \in \Theta$ and all $\delta > 0$, we have*

$$\tilde{\rho}(\theta) - \rho(\theta) \leq \delta + 2\sup_{\theta \in \Theta}\mathbb{P}\left\{|\phi_U(\theta) - 1| \geq \frac{\delta}{2}\right\}\,.$$

Proof The proof is identical to proof of Lemma 3.2 in Medina-Aguayo et al. (2016) by noting that Lemma 3.1 in the same reference holds for two random variables $\phi_U(\theta)$ and $\phi_U(\theta')$ that are not independent, i.e for all $(\theta, \theta') \in \Theta^2$ any $U \in \mathsf{U}_n$ and all $\delta \in (0, 1)$

$$\mathbb{P}\left\{\frac{\phi_U(\theta)}{\phi_U(\theta')} \leq 1 - \delta\right\} \leq 2\sup_{\theta \in \Theta}\mathbb{P}\left\{|\phi_U(\theta) - 1| \geq \delta/2\right\}\,.$$

Lemma 3 *Assume that Assumption* ***A.**4* *holds. Then we have*

$$\sup_{(\theta,\theta')\in\Theta^2} 1 \vee \mathbb{E}\left\{\frac{\phi_U(\theta)}{\phi_U(\theta')}\right\} \leq \mathbb{E}\left\{e^{2\gamma\|\Delta_n(u)\|}\right\} .$$

Proof Using Cauchy-Schwartz inequality, we write that for all $(\theta,\theta')\in\Theta^2$,

$$\begin{aligned}\mathbb{E}\left\{\frac{\phi_U(\theta')}{\phi_U(\theta)}\right\} &= \mathbb{E}\left\{\frac{f(Y_U\,|\,\theta')^{N/n}}{f(Y\,|\,\theta')}\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\right\}\\ &\leq \left[\mathbb{E}\left\{\frac{f(Y_U\,|\,\theta')^{N/n}}{f(Y\,|\,\theta')}\right\}^2\right]^{1/2}\\ &\quad\left[\mathbb{E}\left\{\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\right\}^2\right]^{1/2} . \qquad\qquad (A.8)\end{aligned}$$

Now for all $\theta\in\Theta$, we define the event $\mathcal{E}_\theta := \{U \in \mathsf{U}_n\,,\ f(Y\,|\,\theta) \leq f(Y_U\,|\,\theta)^{N/n}\}$ so that

$$\begin{aligned}\mathbb{E}\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)}\right\}^2 &= \mathbb{E}\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)}\mathbb{1}_{\mathcal{E}_\theta}(U)\right\}^2\\ &\quad+\mathbb{E}\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)}\mathbb{1}_{\overline{\mathcal{E}_\theta}}(U)\right\}^2\end{aligned}$$

and we note that for all $(\theta,U)\in\Theta\times\mathsf{U}_n$, Eq. (26) writes

$$\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)}\right\}^2\mathbb{1}_{\mathcal{E}_\theta}(U) \leq e^{2\gamma\|\Delta_n(U)\|}\mathbb{1}_{\mathcal{E}_\theta}(U)\,,$$

but also

$$\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)}\right\}^2\mathbb{1}_{\overline{\mathcal{E}_\theta}}(U) \leq e^{2\gamma\|\Delta_n(U)\|}\mathbb{1}_{\overline{\mathcal{E}_\theta}}(U)\,,$$

so that

$$\begin{aligned}&\mathbb{E}\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)}\mathbb{1}_{\mathcal{E}_\theta}(U)\right\}^2 + \mathbb{E}\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)}\mathbb{1}_{\overline{\mathcal{E}_\theta}}(U)\right\}^2\\ &\leq \mathbb{E}\left\{e^{2\gamma\|\Delta_n(U)\|}\mathbb{1}_{\mathcal{E}_\theta}(U)\right\} + \mathbb{E}\left\{e^{2\gamma\|\Delta_n(U)\|}\mathbb{1}_{\overline{\mathcal{E}_\theta}}(U)\right\}\\ &= \mathbb{E}\left\{e^{2\gamma\|\Delta_n(U)\|}\right\} .\end{aligned}$$

A similar argument gives the same upper bound for $\mathbb{E}\{f(Y\,|\theta)/f(Y_U\,|\,\theta)^{N/n}\}^2$ so that Eq. (A.8) yields

$$\mathbb{E}\left\{\frac{\phi_U(\theta')}{\phi_U(\theta)}\right\} \leq \mathbb{E}\left\{e^{2\gamma\|\Delta_n(U)\|}\right\} .$$

The proof is completed by noting that for three numbers a, b and c, $c > b \Rightarrow a \vee b \leq a \vee c$ and $\gamma\|\Delta_n(U)\| > 0$. □

Lemma 4 *Assume that Assumption* ***A.**4* *holds. Then we have for all* $\theta\in\Theta$ *and* $\delta > 0$

$$\mathbb{P}\{|\phi_U(\theta)-1| \geq \delta/2\} \leq \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\{\|\Delta_n(U)\|\} .$$

Proof With the same notations as in proof of Lemma 3 and roughly with the same reasoning we have for all $\theta\in\Theta$ and all $\delta > 0$

$$\begin{aligned}&\mathbb{P}\{|\phi_U(\theta)-1| \geq \delta/2\}\\ &\quad= \mathbb{P}\{|\phi_U(\theta)-1| \geq \delta/2 \cap \mathcal{E}_\theta\}\\ &\qquad+\mathbb{P}\left\{|\phi_U(\theta)-1| \geq \delta/2 \cap \overline{\mathcal{E}_\theta}\right\}\\ &\quad= \mathbb{P}\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)} \geq 1+\delta/2 \cap \mathcal{E}_\theta\right\}\\ &\qquad+\mathbb{P}\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)} \leq 1-\delta/2 \cap \overline{\mathcal{E}_\theta}\right\}\\ &\quad\leq \mathbb{P}\left\{e^{\gamma\|\Delta_n(U)\|} \geq 1+\delta/2 \cap \mathcal{E}_\theta\right\}\\ &\qquad+\mathbb{P}\left\{e^{-\gamma\|\Delta_n(U)\|} \leq 1-\delta/2 \cap \overline{\mathcal{E}_\theta}\right\}\\ &\quad\leq \mathbb{P}\{\gamma\|\Delta_n(U)\| \geq \log(1+\delta/2)\}\\ &\qquad+\mathbb{P}\{\gamma\|\Delta_n(U)\| \geq -\log(1-\delta/2)\}\ ,\end{aligned}$$

where the first inequality follows by inclusion (on $\mathcal{E}_\theta$) of

$$\left\{\frac{f(Y_U\,|\,\theta)^{N/n}}{f(Y\,|\,\theta)} \geq 1+\delta/2\right\} \subset \left\{e^{\gamma\|\Delta_n(U)\|} \geq 1+\delta/2\right\}$$

and similarly for the second term. Now, note that for all $x > 0$, $\log(1+x) < -\log(1-x)$ so that

$$\begin{aligned}\mathbb{P}\{|\phi_U(\theta)-1| \geq \delta/2\} &\leq 2\mathbb{P}\{\gamma\|\Delta_n(U)\| \geq \log(1+\delta/2)\}\\ &\leq \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\{\|\Delta_n(U)\|\}\ ,\end{aligned}$$

where the last inequality follows from Markov inequality. □

We study the limiting case where N is fixed and $n \to N$.

Lemma 5 *Assume* N *is fixed and let* $n \to N$*. Then,*

$$\mathbb{E}\{\|\Delta_n(U)\|\} \to 0 \quad \textit{and} \quad \mathbb{E}\{\exp 2\gamma\|\Delta_n(U)\|\} \to 1 .$$

Proof It follows from the fact that when $n \to N$, $\nu_{n,\epsilon}$ converges to the dirac on $U^\dagger = \{1,\ldots,N\}$ and therefore,

$$\begin{aligned}&\mathbb{E}\{\|\Delta_n(U)\|\} \to \|\Delta\bar{S}(U^\dagger)\| = 0 \quad\text{and}\\ &\quad\mathbb{E}\{\exp 2\gamma\|\Delta_n(U)\|\} \to \exp 2\gamma\|\Delta\bar{S}(U^\dagger)\| = 1 .\end{aligned}$$

□

We can now prove Proposition 4:

Proposition *Assume that* ***A.****3 and* ***A.****4 hold. If the marginal MH chain K is geometrically ergodic,* i.e ***A.****1 holds, then there exists an $n_0 \leq N$ such that for all $n > n_0$, $\tilde{K}_n$ is also geometrically ergodic.*

Proof By (Meyn and Tweedie 2009, Theorems 14.0.1 & 15.0.1), there exists a function $V : \mathsf{X} \to [1, \infty[$, two constants $\lambda \in (0, 1)$ and $b < \infty$ and a small set $S \subset \mathsf{X}$ such that K satisfies a drift condition:

$$KV \leq \lambda V + b\mathbb{1}_S \,. \tag{A.9}$$

We now show how to use the previous Lemmas to establish the geometric ergodicity of $\tilde{K}_n$ for some n sufficiently large. This reasoning is very similar to that presented in (Medina-Aguayo et al. 2016, Theorem 3.2).

$$\begin{aligned}
&(\tilde{K}_n - K)V(\theta)\\
&\quad = \int Q(\theta, \mathrm{d}\theta')\left(\tilde{\alpha}(\theta, \theta') - \alpha(\theta, \theta')\right)V(\theta')\\
&\qquad + (\tilde{\rho}(\theta) - \rho(\theta))\,V(\theta)\\
&\quad \leq \left(\mathbb{E}\left\{e^{2\gamma\|\Delta_n(u)\|}\right\} - 1\right)\int Q(\theta, \mathrm{d}\theta')\alpha(\theta, \theta')V(\theta')\\
&\qquad + \left(\delta + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left\{\|\Delta_n(U)\|\right\}\right)V(\theta)\\
&\quad \leq \left(\mathbb{E}\left\{e^{2\gamma\|\Delta_n(u)\|}\right\} - 1\right)(\lambda V(\theta) + b\mathbb{1}_S(\theta) - \rho(\theta)V(\theta))\\
&\qquad + \left(\delta + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left\{\|\Delta_n(U)\|\right\}\right)V(\theta)\\
&\quad \leq \mathbb{E}\left\{e^{2\gamma\|\Delta_n(u)\|}\right\} b\mathbb{1}_S(\theta)\\
&\qquad + \left(\lambda\left(\mathbb{E}\left\{e^{2\gamma\|\Delta_n(u)\|}\right\} - 1\right) + \delta\right.\\
&\qquad \left. + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left\{\|\Delta_n(U)\|\right\}\right)V(\theta)
\end{aligned} \tag{A.10}$$

Combining Eq. (A.9) with Eq. (A.10), we have that

$$\begin{aligned}
\tilde{K}_nV(\theta) \leq &\left\{1 + \mathbb{E}e^{2\gamma\|\Delta_n(u)\|}\right\} b\mathbb{1}_S(\theta)\\
&+ \left(\lambda\mathbb{E}\left\{e^{2\gamma\|\Delta_n(u)\|}\right\} + \delta\right.\\
&\left. + \frac{2\gamma}{\log(1+\delta/2)}\mathbb{E}\left\{\|\Delta_n(U)\|\right\}\right)V(\theta)
\end{aligned} \tag{A.11}$$

Fix $\epsilon > 0$. From Lemma 5, there exists $(n_1, n_2) \in \mathbb{N}^2$ such that

$$\begin{aligned}
&n \geq n_1 \Rightarrow \mathbb{E}\exp\{2\gamma\|\Delta_n(U)\|\} - 1 \leq \epsilon\,,\\
&n \geq n_2 \Rightarrow \mathbb{E}\|\Delta_n(U)\| \leq \epsilon\log(1+\epsilon/4)/4\gamma\,.
\end{aligned} \tag{A.12}$$

Combining Eqs. (A.10) and (A.12) yields that for all $n \geq n_0 := \max(n_1, n_2)$, we have

$$\begin{aligned}
\tilde{K}_nV(\theta) \leq\, &(\epsilon+1)b\mathbb{1}_S(\theta)\\
&+ V(\theta)\left(\lambda(\epsilon+1) + \delta + \frac{\epsilon\log(1+\epsilon/4)}{2\log(1+\delta/2)}\right).
\end{aligned} \tag{A.13}$$

Taking $\delta = \epsilon/2$ in Eq. (A.13) gives

$$\tilde{K}_nV(\theta) \leq (\epsilon+1)b\mathbb{1}_S(\theta) + V(\theta)\left\{\epsilon\,(\lambda+1) + \lambda\right\}.$$

To show that $\tilde{K}_n$ (for $n > n_0$) satisfies a geometric drift condition, it is sufficient to take $\epsilon < (1-\lambda)/(1+\lambda)$ and to check that S is also small for $\tilde{K}_n$. This is demonstrated exactly as in the proof of Medina-Aguayo et al. (2016, Theorem 3.2). □

Appendix A.5 Proof of Proposition 5

This proof borrows ideas from the perturbation analysis of uniformly ergodic Markov chains. First, note that by straightforward algebra we have that

$$\begin{aligned}
&\|K(\theta,\,\cdot\,) - \tilde{K}(\theta,\,\cdot\,)\|\\
&\quad \leq \int Q(\theta, \mathrm{d}\theta')\mathbb{E}\left|\alpha(\theta, \theta') - \tilde{\alpha}(\theta, \theta' \,|\, U)\right|,\\
&\quad \leq \int Q(\theta, \mathrm{d}\theta')\mathbb{E}\left|a(\theta, \theta') - \tilde{a}(\theta, \theta' \,|\, U)\right|,\\
&\quad = \int Q(\theta, \mathrm{d}\theta')a(\theta, \theta')\mathbb{E}\left|1 - \frac{\phi_U(\theta')}{\phi_U(\theta)}\right|,\\
&\quad = \mathbb{E}\left\{\int Q(\theta, \mathrm{d}\theta')a(\theta, \theta')\,|\phi_U(\theta)\right.\\
&\qquad \left. - \phi_U(\theta')|\,\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\right\},\\
&\quad \leq \mathbb{E}\left\{\sup_{\theta\in\Theta}\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\int Q(\theta, \mathrm{d}\theta')a(\theta, \theta')\right.\\
&\qquad \left.\left|\phi_U(\theta) - \phi_U(\theta')\right|\right\},\\
&\quad \leq \mathbb{E}\left\{\sup_{\theta\in\Theta}\frac{f(Y\,|\,\theta)}{f(Y_U\,|\,\theta)^{N/n}}\right\}\sup_{U\in\mathsf{U}_n}\int Q(\theta, \mathrm{d}\theta')a(\theta, \theta')\\
&\qquad \left|\phi_U(\theta) - \phi_U(\theta')\right|.
\end{aligned} \tag{A.14}$$

Now, under Assumption **A**.2 and using Mitrophanov (2005, Corollary 3.1) we have that for any starting point $\theta_0 \in \Theta$,

$$\begin{aligned}
&\|K^i(\theta_0,\,\cdot\,) - \tilde{K}^i(\theta_0,\,\cdot\,)\|\\
&\quad \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho}\right)\sup_{\theta\in\Theta}\|K(\theta,\,\cdot\,) - \tilde{K}(\theta,\,\cdot\,)\|\,,
\end{aligned} \tag{A.15}$$

where $\lambda = \lceil\log(1/C)/\log\rho\rceil$. Combining Eqs (A.14) and (A.15) leads to Eq. (29) with $\kappa = \lambda + C\rho^\lambda/1-\rho$. Moreover, note that using Eq. (29) we have

$$\begin{aligned}\sup_{\theta\in\Theta}\|\pi-\tilde{K}^i(\theta,\,\cdot\,)\| &\le \sup_{\theta\in\Theta}\|\pi-K^i(\theta,\,\cdot\,)\|\\ &\quad+\sup_{\theta\in\Theta}\|K^i(\theta,\,\cdot\,)-\tilde{K}^i(\theta,\,\cdot\,)\|\,,\\ &\le C\rho^i+\kappa A_n \sup_{(\theta,U)\in\Theta\times\mathsf{U}_n} B_n(\theta,U)\end{aligned}$$

and taking the limit when $i\to\infty$ leads to Eq. (30). Finally, for a large enough n, we know from Proposition 4 that the marginal Markov chain $\{\tilde{\theta}_i\,,i\in\mathbb{N}\}$ produced by ISS-MCMC is geometrically ergodic and we denote by $\tilde{\pi}_n$ its stationary distibution. For such a n, we have for any $\theta_0\in\Theta$

$$\begin{aligned}\|\pi-\tilde{\pi}_n\| &\le \|K^i(\theta_0,\,\cdot\,)-\pi\|+\|\tilde{K}^i(\theta_0,\,\cdot\,)-\tilde{\pi}_n\|\\ &\quad+\|K^i(\theta_0,\,\cdot\,)-\tilde{K}^i(\theta_0,\,\cdot\,)\|\\ &\le \|K^i(\theta_0,\,\cdot\,)-\pi\|+\|\tilde{K}^i(\theta_0,\,\cdot\,)-\tilde{\pi}_n\|\\ &\quad+\kappa A_n \sup_{(\theta,U)\in\Theta\times\mathsf{U}_n} B_n(\theta,U)\end{aligned}$$

and taking the limit as $i\to\infty$ yields Eq. (31).

Appendix A.6 Extension of Proposition 5 beyond the time homogeneous case

We start with the two following remarks relative to the Informed Sub-Sampling Markov chain.

Remark 1 Assume $U_0\sim\nu_{n,\epsilon}$ and $\tilde{\theta}_0\sim\mu$ for some initial distribution μ on (Θ,ϑ). The distribution of U_i given $\tilde{\theta}_i$ is for some $u\in\mathsf{U}_n$,

$$\mathbb{P}(U_i=u\,|\,\tilde{\theta}_i)\propto\sum_{U_0\in\mathsf{U}_n}\int_{\tilde{\theta}_0\in\Theta}\nu_{n,\epsilon}(U_0)\mu(\mathrm{d}\tilde{\theta}_0)\bar{K}^i(\tilde{\theta}_0,U_0;\tilde{\theta}_i,u)\,,$$

where $\bar{K}(\theta,U;\mathrm{d}\theta',U'):=K(\theta,\mathrm{d}\theta'\,|\,U)H(U,U')$ and H is the transition kernel of the Markov chain $\{U_i,\,i\in\mathbb{N}\}$. As a consequence $\mathbb{P}(U_i\in\,\cdot\,|\,\tilde{\theta})$ depends on $\tilde{\theta}$ and i.

Remark 2 The marginal Markov chain $\{\tilde{\theta}_i,\;i\in\mathbb{N}\}$ produced by ISS-MCMC algorithm is time inhomogeneous since for all $A\in\mathcal{X}$,

$$\begin{aligned}\tilde{K}(\theta_{i-1},A)&:=\mathbb{P}(\tilde{\theta}_i\in A\,|\,\tilde{\theta}_{i-1})\\ &=\sum_{u\in\mathsf{U}_n}K(\tilde{\theta}_{i-1},\mathrm{d}\tilde{\theta}_i\,|\,U_i)\mathbb{P}(U_i=u\,|\,\tilde{\theta}_i)\,,\end{aligned}\qquad(\text{A.16})$$

and $\mathbb{P}(U_i=u\,|\,\tilde{\theta}_i)$ depends on i (Remark 1). We thus denote by $\tilde{K}_i$ the marginal transition kernel $\tilde{\theta}_{i-1}\to\tilde{\theta}_i$. However, we observe that if the random variables $\{U_i,\;i\in\mathbb{N}\}$ are *i.i.d.* with distribution $\nu_{n,\epsilon}$, K_i becomes time homogeneous as $\mathbb{P}(U_i=u\,|\,\theta_i)=\nu_{n,\epsilon}(u)$ for all i.

A consequence of Remark 2 is that Mitrophanov (2005, Theorem 3.1) does not hold when Assumption **A**.3 is not satisfied. Indeed, $\{\tilde{\theta}_i,\;i\in\mathbb{N}\}$ is not a time homogeneous Markov chain in this case and we first need to generalize the result from Mitrophanov in order to apply it to our context. This is presented in Lemma 6.

Lemma 6 *Let K be the transition kernel of an uniformly ergodic Markov chain that admits π as stationary distribution. Let $\tilde{K}_i$ be the i-th transition kernel of the ISS-MCMC Markov chain. In particular, let $p_i(\,\cdot\,|\,\theta):=\mathbb{P}(U_i\in\,\cdot\,|\,\theta)$ be the distribution of the random variable U_i, used at iteration i of the noisy Markov chain given θ. We have:*

$$\lim_{i\to\infty}\|\pi-\tilde{\pi}_i\|\le\kappa\sup_{\theta\in\Theta}\sup_{i\in\mathbb{N}}\int\delta_i(\theta,\theta')Q(\theta,d\theta')\,,\qquad(\text{A.17})$$

where $\delta_i:\Theta\times\Theta\to\mathbb{R}^+$ is a function that satisfies

$$\mathbb{E}_i\left\{\left|a(\theta,\theta')-\tilde{a}(\theta,\theta'\,|\,U)\right|\right\}\le\delta_i(\theta,\theta')$$

and the expectation is under $p_i(\,\cdot\,|\,\theta)$.

Proof In addition of the notations of Sect. 4, we define the following quantities for a Markov transition kernel regarded as an operator on $\mathcal{M}$, the space of signed measures on $(\Theta,\mathcal{B}(\Theta))$: $\tau(K):=\sup_{\pi\in\mathcal{M}_{0,1}}\|\pi K\|$ is the ergodicity coefficient of K, $\|K\|:=\sup_{\pi\in\mathcal{M}_1}\|\pi K\|$ is the operator norm of K and $\mathcal{M}_1:=\{\pi\in\mathcal{M},\;\|\pi\|=1\}$ and $\mathcal{M}_{0,1}:=\{\pi\in\mathcal{M}_1,\;\pi(\Theta)=0\}$.

Remarks 1 and 2 explain why, in general, $\{\tilde{\theta}_i,\;i\in\mathbb{N}\}$ is a time-inhomogeneous Markov chain with transition kernel $\{\tilde{K}_i,\;i\in\mathbb{N}\}$. For each $i\in\mathbb{N}$, define π_i as the distribution of θ_i produced by the Metropolis–Hastings algorithm (Algorithm 1) with transition kernel K, referred to as the exact kernel hereafter. Our proof is based on the following identity:

$$\begin{aligned}K^i-\tilde{K}_1\tilde{K}_2\cdots\tilde{K}_i&=(K-\tilde{K}_1)K^{i-1}+\tilde{K}_1(K-\tilde{K}_2)K^{i-2}\\ &\quad+\tilde{K}_1\tilde{K}_2(K-\tilde{K}_3)K^{i-3}+\cdots\\ &\quad+\tilde{K}_1\cdots\tilde{K}_{i-1}(K-\tilde{K}_i)\,,\end{aligned}\qquad(\text{A.18})$$

for each $i\in\mathbb{N}$. Equation (A.18) will help translating the proof of Theorem 3.1 in Mitrophanov (2005) to the time-inhomogeneous setting and in particular, we have for each $i\in\mathbb{N}$:

$$\pi_i-\tilde{\pi}_i=(\pi_0-\tilde{\pi}_0)K^i+\sum_{j=0}^{i-1}\tilde{\pi}_j(K-\tilde{K}_{j+1})K^{i-j-1}\,.\qquad(\text{A.19})$$

Following the proof of Theorem 3.1 in Mitrophanov (2005), we obtain

$$\|\pi_i - \tilde{\pi}_i\| \leq \|\pi_0 - \tilde{\pi}_0\|\tau(K^i) + \sum_{j=0}^{i-1} \|K - \tilde{K}_{i-j}\|\tau(K^j)\,,$$
$$\leq \begin{cases} \|\pi_0 - \tilde{\pi}_0\| + i\sup_{j\leq i}\|K - \tilde{K}_j\| & \text{if } i \leq \lambda \\ \|\pi_0 - \tilde{\pi}_0\|C\rho^i + \sup_{j\leq i}\|K - \tilde{K}_j\|\left\{\lambda + C\frac{\rho^\lambda - \rho^i}{1-\rho}\right\} & \text{else} \end{cases} \tag{A.20}$$

where $\lambda = \lceil \log_\rho(1/C) \rceil$. Without loss of generality, we take $\pi_0 = \tilde{\pi}_0$ and since $\|\pi - \tilde{\pi}_i\| \leq \|\pi - \pi_i\| + \|\pi_i - \tilde{\pi}_i\|$ we have for all $i > \lambda$ that

$$\|\pi - \tilde{\pi}_i\| \leq \left\{\lambda + C\frac{\rho^\lambda - \rho^i}{1-\rho}\right\} \sup_{j\leq i}\|K - \tilde{K}_j\|\,. \tag{A.21}$$

Taking the limit as $i \to \infty$ leads to

$$\lim_{i\to\infty}\|\pi - \tilde{\pi}_i\| \leq \left\{\lambda + C\frac{\rho^\lambda}{1-\rho}\right\} \sup_{i\in\mathbb{N}}\|K - \tilde{K}_i\|\,. \tag{A.22}$$

Using a similar derivation than in the proof of Corollary 2.3 in Alquier et al. (2016), we obtain

$$\|K - \tilde{K}_i\| \leq \sup_{\theta\in\Theta}\int Q(\theta, \mathrm{d}\theta')\mathbb{E}_i\left|a(\theta,\theta') - \tilde{a}(\theta,\theta' \mid U_i)\right|\,,$$

where the expectation is under $p_i(\,U \mid \theta)$ and which combined with (A.22) leads to

$$\lim_{i\to\infty}\|\pi_i - \tilde{\pi}_i\| \leq \left(\lambda + C\frac{\rho^\lambda}{1-\rho}\right)\sup_{\theta\in\Theta}\sup_{i\in\mathbb{N}}\mathbb{E}_i\left|a(\theta,\theta')\right.$$
$$\left.-\tilde{a}(\theta,\theta' \mid U_i)\right|$$

where the expectation is under $Q(\theta,\cdot)\otimes p_i(\,\cdot \mid \theta)$. Any upper bound $\delta_i(\theta,\theta')$ of the expectation on the right hand side yields (A.17). □

By straightforward algebra, we have:

$$\mathbb{E}_i\left|a(\theta,\theta') - \tilde{a}(\theta,\theta' \mid U_i)\right|$$
$$= a(\theta,\theta')\mathbb{E}_i\left\{\frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}}\left|\phi_U(\theta) - \phi_U(\theta')\right|\right\} \tag{A.23}$$

where we have defined $\phi_U(\theta) = f(Y_U \mid \theta)^{N/n}/f(Y \mid \theta)$. Using Lemma 6, we have that

$$\lim_{i\to\infty}\|\pi - \tilde{\pi}_i\| \leq \kappa\sup_{\theta\in\Theta}\sup_{i\in\mathbb{N}}\mathbb{E}_i\left\{\sup_{\theta\in\Theta}\frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}}\right.$$
$$\left.\int Q(\theta,\mathrm{d}\theta')a(\theta,\theta')\left|\phi_U(\theta) - \phi_U(\theta')\right|\right\}\,,$$
$$\leq \kappa\sup_{\theta\in\Theta}\sup_{i\in\mathbb{N}}\mathbb{E}_i\left\{\sup_{\theta\in\Theta}\frac{f(Y \mid \theta)}{f(Y_U \mid \theta)^{N/n}}\right\}\sup_{(\theta,U)\in\Theta\times\mathsf{U}_n}$$
$$\int Q(\theta,\mathrm{d}\theta')a(\theta,\theta')\left|\phi_U(\theta) - \phi_U(\theta')\right|\,. \tag{A.24}$$

which is the counterpart of (30) when Assumption **A**.3 does not hold. We note that the second supremum in Eq. (A.24) is in fact B_n defined at Eq. (28) and, as such, can be controlled as described in Section 5.3.1. However, this is not clearly the case for the first supremum in Eq. (A.24) which differs from A_n defined at Eq. (27):

$$\tilde{A}_n := \sup_i\sup_\theta\mathbb{E}_i\{\sup_\theta 1/\phi_U(\theta)\} \neq \mathbb{E}\{\sup_\theta 1/\phi_U(\theta)\} = A_n\,. \tag{A.25}$$

We now show that, under two additional Assumptions (**A**.5 and **A**.6), the control based on the summary statistics also applies to the time inhomogeneous case when Assumption **A**.3 does not hold.

A 5 **One-step minorization** For all $i \in \mathbb{N}$ and all $A \in \vartheta$, there exists some $\eta > 0$ such that $p_i(A) > \eta\lambda(A)$ where λ is the Lebesgue measure.

This assumption typically holds if Θ is compact or if the chain $\{\tilde{\theta}_i, U_i\}_i$ admits a minorization condition. Since we assume, in this discussion, that the exact MH Markov chain is uniformly ergodic and as such satisfy a minorization condition, see e.g. Meyn and Tweedie (2009, Thm 16.2.3) and Hobert and Robert (2004). We may study conditions on which $\{\tilde{\theta}_i\}_i$ inherits this property and leave this for future work but already note that Assumption **A**.5 is not totally unrealistic.

A 6 The marginal Markov chain $\{U_i\}_i$ has initial distribution $U_0 \sim \nu_{n,\epsilon}$.

Even though this assumption is difficult to meet in practice as $|\mathsf{U}_n|$ may be very large, the discussion at the beginning of Section 6.1 indicates an approach to set the distribution of U_0 close from $\nu_{n,\epsilon}$.

Again, while the Assumptions 5 and 6 are perhaps challenging to guarantee, Proposition 8 aims at giving some level of confidence to the user that the ISS-MCMC method is useful, even when Assumption **A**.3 does not hold. In addition, it reinforces the importance of choosing summary statistics that satisfy Assumption **A**.4.

Proposition 8 *Assume that Assumptions **A**.1, **A**.4, **A**.5 and **A**.6 hold. Then there exists a positive number $M > 0$ such that*

$$\tilde{A}_n \leq MA_n\,, \tag{A.26}$$

where A_n and $\tilde{A}_n$ have been defined at Eq. (A.25).

Corollary 2 *Under the same Assumptions as Proposition 8, the control explained in Sect. 5.3.2 is also valid in the time inhomogeneous case.*

Proof of Proposition 8 From Assumption **A.**4, there exists some $\gamma > 0$ such that

$$\tilde{A}_n = \sup_i \sup_\theta \mathbb{E}_i\{f(Y \mid \theta)/f(Y_U \mid \theta)^{N/n}\}$$
$$\leq \sup_i \sup_\theta \int \mathrm{d}p_i(U \mid \theta) e^{\gamma \|\Delta_n(U)\|}\,, \tag{A.27}$$

where $\mathrm{d}p_i(U \mid \theta) = p_i(U \mid \theta)\mathrm{d}U$ and $\mathrm{d}U$ is the counting measure. Now, the conditional probability writes:

$$p_i(U \in \cdot \mid \theta) := \mathbb{P}(U_i \in \cdot\,,\ \tilde{\theta}_i \in \mathrm{d}\theta)/\mathbb{P}(\tilde{\theta}_i \in \mathrm{d}\theta)\,.$$

On the one hand, Lemma 1 shows that there exists a bounded function f_i such that $\mathbb{P}(U_i \in \cdot\,,\ \tilde{\theta}_i \in \mathrm{d}\theta) \leq f_i(\tilde{\theta})\mathrm{d}\theta\nu_{n,\epsilon}(\,\cdot\,)$. On the other hand, Assumption 5 guarantees that there exists some $\eta > 0$ such that for all $\tilde{\theta} \in \Theta$, $\mathbb{P}(\tilde{\theta}_i \in \mathrm{d}\theta) > \eta\mathrm{d}\theta$. Combining those two facts allows to write that

$$p_i(U \in \cdot \mid \theta) \leq \frac{f_i(\theta)\mathrm{d}\theta\nu_{n,\epsilon}(\,\cdot\,)}{\eta\mathrm{d}\theta} = \frac{f_i(\theta)}{\eta}\nu_{n,\epsilon}(\,\cdot\,)\,. \tag{A.28}$$

Plugging Eq. (A.28) into Eq. (A.27), yields to

$$\tilde{A}_n \leq \sup_i \sup_\theta \int \mathrm{d}p_i(U \mid \theta) e^{\gamma \|\Delta_n(U)\|} \leq \sup_\theta \sup_i \frac{f_i(\tilde{\theta})}{\eta} A_n\,,$$

which completes the proof, setting $M := \sup_\theta \sup_i f_i(\theta)/\eta$. □

Lemma 1 *Assume that Assumptions **A.**1, **A.**4, **A.**5 and **A.**6 hold. In addition, let us assume that $U_0 \sim \nu_{n,\epsilon}$. Then $p_i(\theta, U)$ is dominated by $d\theta dU$ where $d\theta$ and dU implicitly refer to the Lebesgue and the counting measure, respectively. In other words there is a sequence of bounded functions $\{f_i : \Theta \to \mathbb{R}^+\}$ such that*

$$dp_i(U, \tilde{\theta}) \leq f_i(\theta)d\tilde{\theta}d\nu_{n,\epsilon}(U)\,. \tag{A.29}$$

Proof We proceed by induction. Defining $\varrho(\tilde{\theta} \mid U)$ as the probability to reject a MH move for the parameter $\tilde{\theta}$ when the subset variable is U, we recall that $\varrho(\tilde{\theta} \mid U) < 1$ and $\tilde{\alpha}(\tilde{\theta}, \tilde{\theta}' \mid U) < 1$. By assumption on the proposal kernel, it satisfies $Q(\tilde{\theta}, \mathrm{d}\tilde{\theta}') = Q(\tilde{\theta}, \tilde{\theta}')\mathrm{d}\tilde{\theta}'$ and define the function $\overline{Q} : \theta \mapsto \sup_{\tilde{\theta}' \in \Theta} Q(\tilde{\theta}', \theta)$. Similarly, we define the function $\overline{\varrho} : \theta \mapsto \sup_{U \in \mathsf{U}_n} \varrho(\theta \mid U)$. Deriving the calculation separately for the continuous and the diagonal parts of the Metropolis–Hastings kernel $K(\theta, \cdot \mid U)$ (see Eq. (25)), we have:

$$\begin{aligned}
&\mathrm{d}p_1(U, \tilde{\theta})\\
&= \int_{\tilde{\theta}_0 \in \Theta} \sum_{U_0 \in \mathsf{U}_n} \mu(\mathrm{d}\tilde{\theta}_0)\nu(U_0)H(U_0, U)K(\tilde{\theta}_0, \mathrm{d}\tilde{\theta} \mid U)\,,\\
&\leq \int \sum \mu(\mathrm{d}\tilde{\theta}_0)\nu(U_0)H(U_0, U)Q(\tilde{\theta}_0, \mathrm{d}\tilde{\theta})\tilde{\alpha}(\tilde{\theta}_0, \tilde{\theta} \mid U)\\
&\quad + \int \sum \mu(\mathrm{d}\tilde{\theta}_0)\nu(U_0)H(U_0, U)\delta_{\tilde{\theta}_0}(\mathrm{d}\tilde{\theta})\varrho(\tilde{\theta}_0 \mid U)\,,\\
&\leq \int \sum \mu(\mathrm{d}\tilde{\theta}_0)\nu(U_0)H(U_0, U)\overline{Q}(\tilde{\theta})\mathrm{d}\tilde{\theta}\\
&\quad + \int \sum \mu(\mathrm{d}\tilde{\theta})\nu(U_0)H(U_0, U)\overline{\varrho}(\tilde{\theta})\,,\\
&\leq \sum \nu(U_0)H(U_0, U)\overline{Q}(\tilde{\theta})\mathrm{d}\tilde{\theta}\\
&\quad + \sum \nu(U_0)H(U_0, U)\mu(\tilde{\theta})\overline{\varrho}(\tilde{\theta})\mathrm{d}\tilde{\theta}\,,\\
&= \underbrace{\left\{\overline{Q}(\tilde{\theta}) + \mu(\tilde{\theta})\overline{\varrho}(\tilde{\theta})\right\}}_{:=f_1(\tilde{\theta})} \mathrm{d}\tilde{\theta}\mathrm{d}\nu(U)\,,
\end{aligned}$$

where the last equality follows from the $\nu_{n,\epsilon}$-stationarity of H. In this derivation, we have defined μ as the initial distribution of the Markov chain $\{\tilde{\theta}_i\}_i$ and ν as a shorthand notation for $\nu_{n,\epsilon}$. Now, let us assume that there is a bounded function f_{i-1} such that $\mathrm{d}p_1(U, \tilde{\theta}) \leq f_{i-1}(\tilde{\theta})\mathrm{d}\tilde{\theta}\mathrm{d}\nu(U)$. Using the notation $\mu K := \int \mu(\mathrm{d}x)K(x, \cdot)$ for any Markov kernel K and a measure μ on some measurable space $(\mathsf{X}, \mathcal{X})$ and recalling that $\bar{K}$ is the transition kernel of ISS-MCMC on the extended space $\Theta \times \mathsf{U}_n$, we have:

$$\begin{aligned}
&\mathrm{d}p_i(U, \tilde{\theta})\\
&= \sum_{U_{i-1} \in \mathsf{U}_n} \int_{\tilde{\theta}_{i-1} \in \Theta} \bar{\mu}\bar{K}^{i-1}(U_{i-1}, \mathrm{d}\tilde{\theta}_{i-1})H(U_{i-1}, U)\\
&\quad K(\tilde{\theta}_{i-1}, \mathrm{d}\tilde{\theta} \mid U)\,,\\
&\leq \sum_{U_{i-1} \in \mathsf{U}_n} \int_{\tilde{\theta}_{i-1} \in \Theta} \bar{\mu}\bar{K}^{i-1}(U_{i-1}, \mathrm{d}\tilde{\theta}_{i-1})H(U_{i-1}, U)\overline{Q}(\tilde{\theta})\mathrm{d}\tilde{\theta}\\
&\quad + \sum_{U_{i-1} \in \mathsf{U}_n} \bar{\mu}\bar{K}^{i-1}(U_{i-1}, \mathrm{d}\tilde{\theta})H(U_{i-1}, U)\varrho(\tilde{\theta} \mid U)\,,\\
&\leq \sum_{U_{i-1} \in \mathsf{U}_n} \bar{\mu}\bar{K}^{i-1}(U_{i-1})H(U_{i-1}, U)\overline{Q}(\tilde{\theta})\mathrm{d}\tilde{\theta}\\
&\quad + \sum_{U_{i-1} \in \mathsf{U}_n} \mathrm{d}p_{i-1}(U_{i-1}, \tilde{\theta})H(U_{i-1}, U)\varrho(\tilde{\theta} \mid U)\\
&\leq \nu(U)\overline{Q}(\tilde{\theta})\mathrm{d}\tilde{\theta} + f_{i-1}(\tilde{\theta}) \sum_{U_{i-1}} H(U_{i-1}, U)\varrho(\tilde{\theta} \mid U)\\
&\leq \underbrace{\left\{\overline{Q}(\tilde{\theta}) + f_{i-1}(\tilde{\theta})\overline{\varrho}(\tilde{\theta})\right\}}_{:=f_i(\tilde{\theta})} \mathrm{d}\tilde{\theta}\mathrm{d}\nu(U)
\end{aligned}$$

and f_i is bounded. The first term in the third inequality follows from noting that

$$\sum \bar{\mu}\bar{K}^{i-1}(U_{i-1})H(U_{i-1}, U_i)$$

$$
\begin{aligned}
&= \sum \int \bar{\mu}\bar{K}^{i-2}(U_{i-2}, \mathrm{d}\tilde{\theta}_{i-2}) \sum \int H(U_{i-2}, U_{i-1}) \\
&\quad K(\tilde{\theta}_{i-2}, \mathrm{d}\tilde{\theta}_{i-1} \mid U_{i-1}) H(U_{i-1}, U) \\
&= \sum \int \bar{\mu}\bar{K}^{i-2}(U_{i-2}, \mathrm{d}\tilde{\theta}_{i-2}) H^2(U_{i-2}, U) = \cdots \\
&= \sum \int \mu(\mathrm{d}\theta_0)\nu(U_0) H^i(U_0, U) = \nu(U)\,.
\end{aligned}
$$

□

Appendix A.7 Proof of Proposition 6

Proof Note that for all $(\theta, \zeta) \in \Theta \times \mathbb{R}^d$, a Taylor expansion of $\pi(\theta)$ and $\phi_U(\theta)$ at $\theta + \Sigma\zeta$ in (32) combined to the triangle inequality leads to:

$$
\begin{aligned}
&B(U, \theta) \\
&\leq \frac{1}{\sqrt{N}} \mathbb{E}\left\{ \left| (M\zeta)^T \nabla_\theta \phi_U(\theta) \right| \left(1 + \frac{1}{\sqrt{N}} (M\zeta)^T \nabla_\theta \log \pi(\theta) \right) \right\} \\
&\quad + \frac{1}{2N} \mathbb{E}\left\{ |(M\zeta)^T \nabla_\theta^2 \phi_U(\theta) M\zeta| \right\} + \mathbb{E}\{R(\|M\zeta\|/\sqrt{N})\}\,,
\end{aligned}
$$

where the expectation is under Φ_d and $R(x) = o(x)$ at 0. Applying Cauchy-Schwartz gives:

$$
\begin{aligned}
B(U, \theta) \leq &\frac{1}{\sqrt{N}} \mathbb{E}\{\|M\zeta\|\} \|\nabla_\theta \phi_U(\theta)\| \\
&+ \frac{1}{N} \mathbb{E}\{\|M\zeta\|^2\} \|\nabla_\theta \phi_U(\theta)\| \|\nabla_\theta \log \pi(\theta)\| \\
&+ \frac{1}{2N} \mathbb{E}\{|\zeta^T M^T \nabla_\theta^2 \phi_U(\theta) M\zeta|\} + \mathbb{E}\{R(\|M\zeta\|/\sqrt{N})\}\,.
\end{aligned}
$$

Now, we observe that:

- $\mathbb{E}\{\|M\zeta\|\} = \mathbb{E}\{\sum_{i=1}^d (\sum_{j=1}^d M_{i,j}\zeta_j)^2\}^{1/2} \leq \mathbb{E}\{\sum_{i=1}^d |\sum_{j=1}^d M_{i,j}\zeta_j|\} \leq \mathbb{E}\{\sum_{i=1}^d \sum_{j=1}^d |M_{i,j}||\zeta_j|\} = \sum_{i=1}^d \sum_{j=1}^d |M_{i,j}| \mathbb{E}\{|\zeta_i|\} = \sqrt{\frac{2}{\pi}} \|M\|_1$
- $\mathbb{E}\{\|M\zeta\|^2\} = \mathbb{E}\{\sum_{i=1}^d (\sum_{j=1}^d M_{i,j}\zeta_j)^2\} = \sum_{i=1}^d \mathbb{E}\{(\sum_{j=1}^d M_{i,j}\zeta_j)^2\} = \sum_{i=1}^d \mathrm{var}(\sum_{j=1}^d M_{i,j}\zeta_j) = \sum_{i=1}^d \sum_{j=1}^d M_{i,j}^2 \mathrm{var}(\zeta_j) = \|M\|_2^2$
- considering the quadratic form associate to the operator $T(U, \theta) = M^T \nabla_\theta^2 \phi_U(\theta) M$, noting that $T(U, \theta)$ is symmetric its eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are real and we have

  $$\zeta^T T(U, \theta)\zeta \leq \lambda_1 \|\zeta\|^2$$

  so that:

  $$
  \begin{aligned}
  &\mathbb{E}\left\{ |(M\zeta)^T \nabla_\theta^2 \phi_U(\theta) M\zeta| \right\} \\
  &\quad \leq d \sup_i |\lambda_i| \leq d \|M^T \nabla_\theta^2 \phi_U(\theta) M\|
  \end{aligned}
  $$

  where for any square matrix A, we have defined $\|A\| = \sup_{x \in \mathbb{R}^d, \|x\|=1} \|Ax\|$ as the operator norm.

□

Appendix B Proof of Proposition 7

In this section, we are assuming that there is an infinite stream of observations $(Y_1, Y_2, \ldots)$ and a parameter $\theta_0 \in \Theta$ such that $Y_i \sim f(\,\cdot \mid \theta_0)$. Let $\rho > 1$ be a constant defined as the ratio N/n i.e the size of the full dataset over the size of the subsamples of interest. The full dataset is thus $Y_{1:\rho n}$. We define the set

$$\mathsf{U}_n^\rho = \{U \subset \{1, \ldots, \rho n\},\ |U| = n\}$$

such that Y_U $(U \in \mathsf{U}_n^\rho)$ is the set of subsamples of interest. We study the asymptotics when $n \to \infty$ i.e we let the whole dataset and the size of subsamples of interest grow at the same rate.

Proposition 7 *Let $\theta_{\rho n}^*$ be the MLE of $Y_1, \ldots, Y_{\rho n}$ and θ_U^* be the MLE of the subsample Y_U $(U \in \mathsf{U}_n^\rho)$. Assume that there exists a compact set $\kappa_n \subset \Theta$ such that $(\theta_{\rho n}^*, \theta_0) \in \kappa_n^2$ and for all U, there exists a compact set $\kappa_U \subset \Theta$ such that $(\theta_U^*, \theta_0) \in \kappa_U^2$. Then, there exists a constants β, a metric $\|\cdot\|_{\theta_0}$ on Θ and a non-decreasing subsequence $\{\sigma_n\}_{n \in \mathbb{N}}$, $(\sigma_n \in \mathbb{N})$ such that for all $U \in \mathsf{U}_{\sigma_n}^\rho$, we have for p-almost all $\theta \in \kappa_n \cap \kappa_U$*

$$
\begin{aligned}
&\log f(Y_{1:\rho\sigma_n} \mid \theta) - \rho \log f(Y_U \mid \theta) \leq H_n(Y, \theta) + \beta \\
&\quad + \frac{\rho\sigma_n}{2} \|\theta_U^* - \theta^*\|_{\theta_0}\,, \qquad \text{(B.1)}
\end{aligned}
$$

where

$$\underset{n \to \infty}{plim} \quad H_n(Y, \theta) \overset{\mathbb{P}_{\theta_0}}{=} 0\,.$$

Proof Fix $n \in \mathbb{N}$. Consider the case where the prior distribution p is uniform on κ_n. In this case, the posterior is

$$
\begin{aligned}
&\pi_n(\theta \mid Y_{1:\rho n}) = f(Y_{1:\rho n} \mid \theta) \mathbb{1}_{\kappa_n}(\theta) \big/ Z_{\rho n}\,, \\
&\qquad Z_{\rho n} = \int_{\kappa_n} f(Y_{1:\rho n} \mid \theta) \mathrm{d}\theta
\end{aligned}
$$

and from Corollary 3, we know that there exists a subsequence $\tau_n \subset \mathbb{N}$ such that for p-almost all $\theta \in \kappa_n$

$$\left| \log \frac{f(Y_{1:\rho\tau_n} \mid \theta)}{Z_{\rho\tau_n}} - \log \Phi_{\rho\tau_n}(\theta) \right| \overset{\mathbb{P}_{\theta_0}}{\to} 0\,, \qquad \text{(B.2)}$$

where $\theta \mapsto \Phi_{\rho\tau_n}(\theta)$ is the pdf of $\mathcal{N}(\theta_{\rho\tau_n}^*, I(\theta_0)^{-1}/\rho\tau_n)$. Similarly, there exists another subsequence $\gamma_n \subset \mathbb{N}$ such that for all $U \in \mathsf{U}^{\rho}_{\gamma_n}$ and for p-almost all $\theta \in \kappa_U$

$$\begin{aligned}&\left|\rho \log \frac{f(Y_U\,|\,\theta)}{Z_{\gamma_n}(U)} - \rho \log \Phi_U(\theta)\right| \overset{\mathbb{P}_{\theta_0}}{\to} 0\,,\\ &Z_{\gamma_n}(U) = \int_{\kappa_U} f(Y_U\,|\,\theta)\mathrm{d}\theta\end{aligned} \tag{B.3}$$

where $\theta \mapsto \Phi_U(\theta)$ is the pdf of $\mathcal{N}(\theta^*_U, I(\theta_0)^{-1}/|U|)$. Let $\{\sigma_n\}_{n\in\mathbb{N}}$ be the sequence defined as $\sigma_n = \max\{\tau_n, \gamma_n\}$. We know from (B.2) and (B.3) that for all $\varepsilon > 0$ and all $\eta > 0$, there exists $n_1 \in \mathbb{N}$ such that for all $U \in \mathsf{U}^{\rho}_{\sigma_n}$ and for all $n \geq n_1$

$$\begin{aligned}\mathbb{P}_{\theta_0}&\left\{\left|\log \frac{f(Y_{1:\rho\sigma_n}\,|\,\theta)}{Z_{\rho\sigma_n}} - \log \Phi_{\rho\sigma_n}(\theta)\right|\right.\\ &\left.+\left|\rho \log \frac{f(Y_U\,|\,\theta)}{Z_{\sigma_n}(U)} - \rho \log \Phi_U(\theta)\right| \geq \varepsilon\right\} \leq \eta\,.\end{aligned} \tag{B.4}$$

Now, by straightforward algebra, we have for any $U \in \mathsf{U}^{\rho}_{\sigma_n}$

$$\begin{aligned}&\log f(Y_{1:\rho\sigma_n}\,|\,\theta) - \rho \log f(Y_U\,|\,\theta)\\ &\quad= \log \frac{f(Y_{1:\rho\sigma_n}\,|\,\theta)}{Z_{\rho\sigma_n}} - \log \Phi_{\rho\sigma_n}(\theta)\\ &\qquad - \rho \log \frac{f(Y_U\,|\,\theta)}{Z_{\sigma_n}(U)}\\ &\qquad + \rho \log \Phi_U(\theta) + \log \frac{Z_{\rho\sigma_n}}{Z_{\sigma_n}(U)^{\rho}} + \log \Phi_{\rho\sigma_n}(\theta)\\ &\qquad - \rho \log \Phi_U(\theta)\\ &\quad\leq \left|\log \frac{f(Y_{1:\rho\sigma_n}\,|\,\theta)}{Z_{\rho\sigma_n}} - \log \Phi_{\rho\sigma_n}(\theta)\right.\\ &\qquad \left.-\rho \log \frac{f(Y_U\,|\,\theta)}{Z_{\sigma_n}(U)} + \rho \log \Phi_U(\theta)\right|\\ &\qquad + \log \frac{Z_{\rho\sigma_n}}{Z_{\sigma_n}(U)^{\rho}} + (\rho - 1)\log(2\pi)^{d/2}\\ &\qquad + \frac{\rho\sigma_n}{2}\Big|\|\theta - \theta^*_U\|_{\theta_0} - \|\theta - \theta^*\|_{\theta_0}\Big|\\ &\quad\leq \left|\log \frac{f(Y_{1:\rho\sigma_n}\,|\,\theta)}{Z_{\rho\sigma_n}} - \log \Phi_{\rho\sigma_n}(\theta)\right|\\ &\qquad + \left|\rho \log \frac{f(Y_U\,|\,\theta)}{Z_{\sigma_n}(U)} - \rho \log \Phi_U(\theta)\right|\\ &\qquad + \log \frac{Z_{\rho\sigma_n}}{Z_{\sigma_n}(U)^{\rho}} + (\rho - 1)\log(2\pi)^{d/2}\\ &\qquad + \frac{\rho\sigma_n}{2}\|\theta^*_U - \theta^*\|_{\theta_0}\,,\end{aligned} \tag{B.5}$$

where we have used Lemma 3 for the first inequality and the triangle inequalities for the second. Combining (B.5) with (B.4) yields (34). □

Lemma 2 *Consider a posterior distribution π_n given n data $Y_{1:n}$ where p is the prior distribution and its Bernstein-von Mises approximation is $\Phi_n = \mathcal{N}(\theta^*(Y_{1:n}), I(\theta_0)^{-1}/n)$. There exists a subsequence $\{\tau_n\}_n \subset \mathbb{N}$ such that*

$$\underset{n\to\infty}{\mathit{plim}}\left|\pi_{\tau_n}(\theta) - \Phi_{\tau_n}(\theta)\right| \overset{\mathbb{P}_{\theta_0}}{=} 0\,, \quad \textit{for } p\textit{-almost all } \theta\,. \tag{B.6}$$

Proof This follows for the fact that convergence in L_1 implies pointwise convergence almost everywhere of a subsequence, i.e there exists a subsequence $\{\tau_n\}_{n\in\mathbb{N}} \subset \mathbb{N}$ such that

$$\|\pi_n - \Phi_n\|_1 \to 0 \Rightarrow |\pi_{\tau_n}(\theta) - \Phi_{\tau_n}(\theta)| \to 0 \quad p\text{-a.e.} \tag{B.7}$$

Eq. B.6 follows from combining the Bernstein-von Mises theorem and Eq. (B.7):

$$\begin{aligned}&\mathrm{plim}_{n\to\infty}\|\pi_n - \Phi_n(\theta^*, I(\theta_0)^{-1}/n)\|_1 \overset{\mathbb{P}_{\theta_0}}{=} 0\\ &\quad\Rightarrow \mathrm{plim}_{n\to\infty}|\pi_{\tau_n}(\theta) - \Phi_{\tau_n}(\theta)| \overset{\mathbb{P}_{\theta_0}}{=} 0 \quad p\text{-a.e.}\end{aligned}$$

□

Corollary 3 *There exists a subsequence $\{\tau_n\}_{n\in\mathbb{N}} \in \mathbb{N}$ such that*

$$\underset{n\to\infty}{\mathit{plim}}\left|\log \pi_{\tau_n}(\theta) - \log \Phi_{\tau_n}(\theta)\right| \overset{\mathbb{P}_{\theta_0}}{=} 0\,, \quad \textit{for } p\textit{-almost all } \theta\,. \tag{B.8}$$

Proof Follows from Lemma 2, by continuity of the logarithm.

Lemma 3 *For any $U \in \mathsf{U}_n$, let $\theta \mapsto \Phi_U(\theta)$ be the pdf of $\mathcal{N}(\theta^*_U, I(\theta_0)^{-1}/n)$ and $\Phi_{\rho n}$ be the pdf of $\mathcal{N}(\theta^*_{\rho n}, I(\theta_0)^{-1}/\rho n)$ be the Bernstein-von Mises approximations of respectively $\pi(\cdot\,|\,Y_U)$ and $\pi(\cdot\,|\,Y_{1:\rho n})$ where $U \subset \mathsf{U}_n(Y_{1:\rho n})$. Then we have for all $\theta \in \Theta$*

$$\begin{aligned}&\log \Phi_{\rho n}(\theta) - \rho \log \Phi_U(\theta) \leq (\rho - 1)\log(2\pi)^{d/2}\\ &\quad + \frac{\rho n}{2}\left\{\|\theta - \theta^*_U\|_{\theta_0} - \|\theta - \theta^*\|_{\theta_0}\right\}\,,\end{aligned}$$

where for any d-squared symmetric matrix M, we have defined by $\|\cdot\|_M$ the norm associated to the scalar product $\langle u, v\rangle_M = u^T M v$.

Proof This follows from straightforward algebra and noting that

$$\log \rho n|I(\theta_0)| - \rho \log n|I(\theta_0)| \leq 0\,.$$

□

References

Allassonnière, S., Amit, Y., Trouvé, A.: Towards a coherent statistical framework for dense deformable template estimation. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **69**(1), 3–29 (2007)

Alquier, P., Friel, N., Everitt, R., Boland, A.: Noisy Monte Carlo: convergence of Markov chains with approximate transition kernels. Stat. Comput. **26**(1–2), 29–47 (2016)

Andrieu, C., Roberts, G.O.: The pseudo-marginal approach for efficient Monte Carlo computations. Ann. Stat. **37**, 697–725 (2009)

Andrieu, C., Vihola, M.: Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. Ann Appl. Probab. **25**(2), 1030–1077 (2015)

Banterle, M., Grazian, C., Lee, A., Robert, C.P.: Accelerating Metropolis–Hastings algorithms by delayed acceptance. arXiv preprint arXiv:1503.00996 (2015)

Bardenet, R., Doucet, A., Holmes, C.: Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In: ICML, pp. 405–413 (2014)

Bardenet, R., Doucet, A., Holmes, C.: On Markov chain Monte Carlo methods for tall data. J. Mach. Learn. Res. **18**, 1–43 (2017)

Bierkens, J., Fearnhead, P., Roberts, G.: The zig-zag process and super-efficient sampling for Bayesian analysis of big data. Ann. Stat. (2018) **(to appear)**

Chib, S., Greenberg, E.: Understanding the metropolis-Hastings algorithm. Am. Stat. **49**(4), 327–335 (1995)

Csilléry, K., Blum, M.G., Gaggiotti, O.E., François, O.: Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evolut. **25**(7), 410–418 (2010)

Dalalyan, A.S.: Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. arXiv preprint arXiv:1704.04752 (2017)

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

Fearnhead, P., Bierkens, J., Pollock, M., Roberts, G.O.: Piecewise deterministic Markov processes for continuous-time Monte Carlo. arXiv preprint arXiv:1611.07873 (2016)

Fearnhead, P., Prangle, D.: Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. R. Stat. Soc. Seri. B (Stat. Methodol.) **74**(3), 419–474 (2012)

Geyer, C.J., Thompson, E.A.: Annealing Markov chain Monte Carlo with applications to ancestral inference. J. Am. Stat. Assoc. **90**(431), 909–920 (1995)

Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. Bernoulli **7**, 223–242 (2001)

Hobert, J.P., Robert, C.P.: A mixture representation of π with applications in Markov chain Monte Carlo and perfect sampling. Ann. Appl. Probab. **14**, 1295–1305 (2004)

Huggins, J., Zou, J.: Quantifying the accuracy of approximate diffusions and Markov chains. In: Proceedings of the 20th International Conference on Artifical Intelligence and Statistics, PLMR, vol. 54, pp. 382–391 (2016)

Jacob, P.E., Thiery, A.H., et al.: On nonnegative unbiased estimators. Ann. Stat. **43**(2), 769–784 (2015)

Johndrow, J.E., Mattingly, J.C.: Error bounds for approximations of Markov chains. arXiv preprint arXiv:1711.05382 (2017)

Johndrow, J.E., Mattingly, J.C., Mukherjee, S., Dunson, D.: Approximations of Markov chains and Bayesian inference. arXiv preprint arXiv:1508.03387 (2015)

Korattikara, A., Chen, Y., Welling, M.: Austerity in MCMC land: cutting the Metropolis–Hastings budget. In: Proceedings of the 31st International Conference on Machine Learning (2014)

Le Cam, L.: On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. Univ. Calif. Publ. Stat. **1**, 277–330 (1953)

Le Cam, L.: Asymptotic Methods in Statistical Decision Theory. Springer, Berlin (1986)

Maclaurin, D., Adams, R.P.: Firefly Monte Carlo: exact MCMC with subsets of data. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)

Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. Stat. Comput. **22**(6), 1167–1180 (2012)

Medina-Aguayo, F.J., Lee, A., Roberts, G.O.: Stability of noisy Metropolis-Hastings. Stat. Comput. **26**(6), 1187–1211 (2016)

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)

Meyn, S.P., Tweedie, R.L.: Markov Chains and Stochastic Stability. Cambridge University Press, Cambridge (2009)

Mitrophanov, A.Y.: Sensitivity and convergence of uniformly ergodic Markov chains. J. Appl. Probab. **142**, 003–1014 (2005)

Nunes, M.A., Balding, D.J.: On optimal selection of summary statistics for approximate Bayesian computation. Stat. Appl. Genet. Mol. Biol. **9**(1) (2010)

Pollock, M., Fearnhead, P., Johansen, A.M., Roberts, G.O.: The scalable Langevin exact algorithm: Bayesian inference for big data. arXiv preprint arXiv:1609.03436 (2016)

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol. Biol. Evol. **16**(12), 1791–1798 (1999)

Quiroz, M., Villani, M., Kohn, R.: Speeding up MCMC by efficient data subsampling. Riksbank Research Paper Series (121) (2015)

Quiroz, M., Villani, M., Kohn, R.: Exact subsampling MCMC. arXiv preprint arXiv:1603.08232 (2016)

Roberts, G.O., Rosenthal, J.S., et al.: Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. **16**(4), 351–367 (2001)

Rudolf, D., Schweizer, N.: Perturbation theory for Markov chains via Wasserstein distance. Bernoulli **24**(4A), 2610–2639 (2018)

Van der Vaart, A.W.: Asymptotic Statistics, vol. 3. Cambridge University Press, Cambridge (2000)

Welling, M., Teh, Y.W.: Bayesian learning via stochastic gradient Langevin dynamics. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 681–688 (2011)

Wilkinson, R.D.: Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Stat. Appl. Genet. Mol. Biol **12**(2), 129–141 (2013)