Informed Subsampling MCMC:
Approximate Bayesian Inference for Large Datasets

Pierre Alquier

Bayesian Statistics in the Big Data Era

F. Maire, N. Friel, P. Alquier, Informed Sub-Sampling MCMC : Approximate Bayesian Inference for Large Datasets. To appear in *Statistics and Computing*.

F. Maire, N. Friel, P. Alquier, Informed Sub-Sampling MCMC : Approximate Bayesian Inference for Large Datasets. To appear in *Statistics and Computing*.

F. Maire, N. Friel, P. Alquier, Informed Sub-Sampling MCMC : Approximate Bayesian Inference for Large Datasets. To appear in *Statistics and Computing*.

Metropolis-Hastings algorithm

- Consider the posterior distribution

$$\pi(\theta|Y) = \pi(\theta \mid Y_1, \ldots, Y_N) \propto f(Y_1, \ldots, Y_N \mid \theta)\pi(\theta)$$

Metropolis-Hastings algorithm

- Consider the posterior distribution

$$\pi(\theta|Y) = \pi(\theta \mid Y_1, \ldots, Y_N) \propto f(Y_1, \ldots, Y_N \mid \theta)\pi(\theta)$$

- Metropolis-Hastings, transitioning from $\theta_k \to \theta_{k+1}$:

  1. draw $\tilde{\theta} \sim Q(\theta_k, \cdot)$
  2. set $\theta_{k+1} = \tilde{\theta}$ with probability

$$A(\theta_k, \tilde{\theta}) = 1 \wedge \frac{\pi(\tilde{\theta} \mid Y) Q(\tilde{\theta}, \theta_k)}{\pi(\theta_k \mid Y) Q(\theta_k, \tilde{\theta})}$$

  and $\theta_{k+1} = \theta_k$ w.p. $1 - A(\theta_k, \tilde{\theta})$.

Metropolis-Hastings algorithm

- Consider the posterior distribution

$$\pi(\theta|Y) = \pi(\theta \,|\, Y_1, \ldots, Y_N) \propto f(Y_1, \ldots, Y_N \,|\, \theta)\pi(\theta)$$

- Metropolis-Hastings, transitioning from $\theta_k \to \theta_{k+1}$:

  1. draw $\tilde{\theta} \sim Q(\theta_k, \cdot)$
  2. set $\theta_{k+1} = \tilde{\theta}$ with probability

$$A(\theta_k, \tilde{\theta}) = 1 \wedge \frac{\pi(\tilde{\theta} \,|\, Y) Q(\tilde{\theta}, \theta_k)}{\pi(\theta_k \,|\, Y) Q(\theta_k, \tilde{\theta})}$$

  and $\theta_{k+1} = \theta_k$ w.p. $1 - A(\theta_k, \tilde{\theta})$.

Is it possible to usefully employ MH without computing $\pi(\tilde{\theta} \,|\, Y_1, \ldots, Y_N)$, to accept/reject $\tilde{\theta}$ based on a subset of $n \ll N$ data ?

Austerity in MCMC-Land

A. Korattikara, Y. Chen, M. Welling, Austerity in MCMC land : Cutting the Metropolis-Hastings budget. *ICML* (2014).

Austerity in MCMC-Land

A. Korattikara, Y. Chen, M. Welling, Austerity in MCMC land : Cutting the Metropolis-Hastings budget. *ICML* (2014).

- Idea : $\theta_{k+1} = \tilde{\theta}$ iff

$$U \leq \frac{\pi(\tilde{\theta} \,|\, Y) Q(\tilde{\theta}, \theta_k)}{\pi(\theta_k \,|\, Y) Q(\theta_k, \tilde{\theta})} \text{ where } U \sim \mathcal{U}[0,1].$$

Austerity in MCMC-Land

A. Korattikara, Y. Chen, M. Welling, Austerity in MCMC land : Cutting the Metropolis-Hastings budget. *ICML* (2014).

- Idea : $\theta_{k+1} = \tilde{\theta}$ iff

$$U \leq \frac{\pi(\tilde{\theta} \,|\, Y) Q(\tilde{\theta}, \theta_k)}{\pi(\theta_k \,|\, Y) Q(\theta_k, \tilde{\theta})} \text{ where } U \sim \mathcal{U}[0,1].$$

- In the i.i.d case $f(Y_1, \ldots, Y_N \,|\, \theta) = \prod_{i=1}^N f(Y_i \,|\, \theta)$,

$$\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{f(Y_i|\tilde{\theta})}{f(Y_i|\theta_k)} \right) \geq s := \frac{1}{N} \log \left[U \frac{\pi(\theta_k) Q(\theta_k, \tilde{\theta})}{\pi(\tilde{\theta}) Q(\tilde{\theta}, \theta_k)} \right]$$

Austerity in MCMC-Land

A. Korattikara, Y. Chen, M. Welling, Austerity in MCMC land : Cutting the Metropolis-Hastings budget. *ICML* (2014).

- In the i.i.d case $f(Y_1, \ldots, Y_N \mid \theta) = \prod_{i=1}^{N} f(Y_i \mid \theta)$,

$$\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f(Y_i|\tilde{\theta})}{f(Y_i|\theta_k)}\right) \geq s := \frac{1}{N}\log\left[u\frac{\pi(\theta_k)Q(\theta_k,\tilde{\theta})}{\pi(\tilde{\theta})Q(\tilde{\theta},\theta_k)}\right]$$

Based on a random subsample of size n, test

$$\mathbf{H}_0 : \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f(Y_i|\tilde{\theta})}{f(Y_i|\theta_k)}\right) \geq s.$$

Austerity in MCMC-Land

A. Korattikara, Y. Chen, M. Welling, Austerity in MCMC land : Cutting the Metropolis-Hastings budget. *ICML* (2014).

- In the i.i.d case $f(Y_1, \ldots, Y_N \mid \theta) = \prod_{i=1}^{N} f(Y_i \mid \theta)$,

$$\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f(Y_i|\tilde{\theta})}{f(Y_i|\theta_k)}\right) \geq s := \frac{1}{N}\log\left[U\frac{\pi(\theta_k)Q(\theta_k,\tilde{\theta})}{\pi(\tilde{\theta})Q(\tilde{\theta},\theta_k)}\right]$$

Based on a random subsample of size n, test

$$\mathbf{H}_0 : \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f(Y_i|\tilde{\theta})}{f(Y_i|\theta_k)}\right) \geq s.$$

If the power of the test is not large enough : increase n and start again.

Theoretical analysis

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

Theoretical analysis

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

- Designed an explicit test with probability of mistake $\leq \delta$.

Theoretical analysis

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

- Designed an explicit test with probability of mistake $\leq \delta$.
- Theoretical analysis :

Theorem ($\simeq$ BDH 2014)

Assume that the transition kernel $P(\cdot, \cdot)$ of the (regular) MH algorithm is uniformly geometrically ergodic, that is

$$\exists \rho < 1, \forall \theta, \ \left\| P^k(\theta, \cdot) - \pi(\cdot | Y) \right\|_{\mathrm{TV}} \leq C \rho^k.$$

Theoretical analysis

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

- Designed an explicit test with probability of mistake $\leq \delta$.
- Theoretical analysis :

Theorem ($\simeq$ BDH 2014)

Assume that the transition kernel $P(\cdot, \cdot)$ of the (regular) MH algorithm is uniformly geometrically ergodic, that is

$$\exists \rho < 1, \forall \theta, \ \left\| P^k(\theta, \cdot) - \pi(\cdot | Y) \right\|_{\mathrm{TV}} \leq C \rho^k.$$

Then the kernel $\tilde{P}(\cdot, \cdot)$ of the subsampling MH satisfies

$$\forall \theta, \ \left\| \tilde{P}^k(\theta, \cdot) - P^k(\theta, \cdot) \right\|_{\mathrm{TV}} \leq \tilde{C} \delta.$$

Generalization : Noisy MCMC

P. Alquier, N. Friel, R. G. Everitt & A. Boland. Noisy Monte-Carlo : Convergence of Markov Chains with Approximate Transition Kernels. *Statistics and Computing*, 2016.

Generalization : Noisy MCMC

P. Alquier, N. Friel, R. G. Everitt & A. Boland. Noisy Monte-Carlo : Convergence of Markov Chains with Approximate Transition Kernels. *Statistics and Computing*, 2016.

- Noisy MCMC :
  1. draw $\tilde{\theta} \sim Q(\theta_k, \cdot)$
  2. set $\theta_{k+1} = \tilde{\theta}$ with probability $\hat{A}(\theta_k, \tilde{\theta}, X)$.
- $\hat{A}(\theta, \theta', X)$: approximation of the acceptance ratio $A(\theta, \theta')$ that might be based on the drawing of an auxiliary random variable X.

Generalization : Noisy MCMC

P. Alquier, N. Friel, R. G. Everitt & A. Boland. Noisy Monte-Carlo : Convergence of Markov Chains with Approximate Transition Kernels. *Statistics and Computing*, 2016.

- Noisy MCMC :
  1. draw $\tilde{\theta} \sim Q(\theta_k, \cdot)$
  2. set $\theta_{k+1} = \tilde{\theta}$ with probability $\hat{A}(\theta_k, \tilde{\theta}, X)$.
- $\hat{A}(\theta, \theta', X)$: approximation of the acceptance ratio $A(\theta, \theta')$ that might be based on the drawing of an auxiliary random variable X.
- Define

$$\delta(\theta, \theta') = \mathbb{E}_X |\hat{A}(\theta, \theta', X) - A(\theta, \theta')|$$

$$\delta = \sup_{\theta} \int Q(\theta, \mathrm{d}\theta') \delta(\theta, \theta').$$

Theoretical study of Noisy MCMC

Theorem (AFEB 2016)

Assume that the transition kernel $P(\cdot, \cdot)$ of MH algorithm is uniformly geometrically ergodic with (C, ρ). Then the kernel $\tilde{P}(\cdot, \cdot)$ of noisy MCMC satisfies

$$\forall \theta, \ \left\| \tilde{P}^k(\theta, \cdot) - P^k(\theta, \cdot) \right\|_{\mathrm{TV}} \leq \tilde{C} \delta.$$

Theoretical study of Noisy MCMC

Theorem (AFEB 2016)

Assume that the transition kernel $P(\cdot, \cdot)$ of MH algorithm is uniformly geometrically ergodic with (C, ρ). Then the kernel $\tilde{P}(\cdot, \cdot)$ of noisy MCMC satisfies

$$\forall \theta, \ \left\| \tilde{P}^k(\theta, \cdot) - P^k(\theta, \cdot) \right\|_{\mathrm{TV}} \leq \tilde{C} \delta.$$

Finally,

D. Rudolf, N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 2018.

extended considerably the generality of the results : (non-uniformly) geometrically erdogic Markov chains.

A problem with the "austerity" approach

- remind that in order to reach a proba. of mistake $\leq \delta$, one might have to increase the size of the subsample n, depending on the values of θ_k and $\tilde{\theta}$.

A problem with the "austerity" approach

- remind that in order to reach a proba. of mistake $\leq \delta$, one might have to increase the size of the subsample n, depending on the values of θ_k and $\tilde{\theta}$.
- Bardenet, Doucet & Holmes actually show that in the stationary regime, one has

$$n \propto N$$

for some constant N.

A problem with the "austerity" approach

- remind that in order to reach a proba. of mistake $\leq \delta$, one might have to increase the size of the subsample n, depending on the values of θ_k and $\tilde{\theta}$.
- Bardenet, Doucet & Holmes actually show that in the stationary regime, one has

$$n \propto N$$

  for some constant N.
- But one would like $n \ll N$.

Firefly MCMC

D., Maclaurin, R., Adams. Firefly Monte Carlo : Exact MCMC with Subsets of Data. *IJCAI*, 2014.

Firefly MCMC

D., Maclaurin, R., Adams. Firefly Monte Carlo : Exact MCMC with Subsets of Data. *IJCAI*, 2014.

$$\text{Aux. variables } Z_i = \begin{cases} 1 \text{ if } Y_i \text{ included in the subsample,} \\ 0 \text{ otherwise.} \end{cases}$$

Adapting acceptance ratio $\rightarrow$ exact sampling,

Firefly MCMC

D., Maclaurin, R., Adams. Firefly Monte Carlo : Exact MCMC with Subsets of Data. *IJCAI*, 2014.

$$\text{Aux. variables } Z_i = \begin{cases} 1 \text{ if } Y_i \text{ included in the subsample,} \\ 0 \text{ otherwise.} \end{cases}$$

Adapting acceptance ratio $\rightarrow$ exact sampling, but one has to know an easily computable lower-bound $0 < B_i(\theta) < f(Y_i|\theta)$.

Firefly MCMC

D., Maclaurin, R., Adams. Firefly Monte Carlo : Exact MCMC with Subsets of Data. *IJCAI*, 2014.

$$\text{Aux. variables } Z_i = \begin{cases} 1 \text{ if } Y_i \text{ included in the subsample,} \\ 0 \text{ otherwise.} \end{cases}$$

Adapting acceptance ratio $\rightarrow$ exact sampling, but one has to know an easily computable lower-bound $0 < B_i(\theta) < f(Y_i|\theta)$.

If the lower-bound B is not tight, one ends up with

$$n = \sum_{i=1}^{N} Z_i \simeq N.$$

Subposteriors

Definition

Let Y_U be a subset of $Y_{1:N}$ of size n, we define the "scaled subposteriors" $\bar{\pi}_n$ by

$$\bar{\pi}_n(\theta \mid Y_U) \propto f(Y_U \mid \theta)^{\frac{N}{n}} \pi(\theta) = \left[\prod_{i \in U} f(Y_i|\theta) \right]^{\frac{N}{n}} \pi(\theta)$$

Motivating example : exponential family

Assume that f belongs to the exponential family

$$f(y \mid \theta) \propto \psi(\theta) \exp\{\phi(\theta)^T S(y)\} .$$

Motivating example : exponential family

Assume that f belongs to the exponential family

$$f(y \mid \theta) \propto \psi(\theta) \exp\{\phi(\theta)^T S(y)\}.$$

Definition

For any $U \in \mathcal{U}_n$, the set of all possible subset of $\{1, \ldots, N\}$ of size n, define the vector of sufficient statistics between the whole dataset and the sub-sample Y_U as :

$$\Delta_n(U) = \sum_{k=1}^{N} S(y_k) - \frac{N}{n} \sum_{k \in U} S(y_k).$$

Exponential family : an optimality result

Proposition

For any $U \in \mathcal{U}_n$, the following inequality holds :

$$KL(\pi(\cdot|Y), \bar{\pi}_n(\cdot|Y_U)) \leq B(Y, U) \text{ , where}$$

$$B(Y, U) = \|\Delta_n(U)\| \log \mathbb{E}_\pi \exp\left\{ \|\phi(\theta) - \mathbb{E}_\pi(\phi(\theta))\| \right\} .$$

Exponential family : an optimality result

Proposition

For any $U \in \mathcal{U}_n$, the following inequality holds :

$$KL(\pi(\cdot|Y), \bar{\pi}_n(\cdot|Y_U)) \leq B(Y, U)\,, \text{ where}$$

$$B(Y, U) = \|\Delta_n(U)\| \log \mathbb{E}_\pi \exp\left\{ \|\phi(\theta) - \mathbb{E}_\pi(\phi(\theta))\| \right\}.$$

As a consequence,

1. If there is a "perfect" U, that is

$$\frac{1}{N}\sum_{k=1}^{N} S(y_k) = \frac{1}{n}\sum_{k\in U} S(y_k),$$

then $\bar{\pi}_n(\cdot|Y_U) = \pi(\cdot|Y)$.

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

Toy example : Bernoulli model

Simulate $N = 10,000$ observations $Y_1, \ldots, Y_N \sim \mathcal{B}e(p)$. Then

$$\pi(\theta \,|\, Y) \propto \pi(\theta)(1-p)^{\sum_{k=1}^{N} Y_k} p^{N - \sum_{k=1}^{N} Y_k}$$

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

Toy example : Bernoulli model

Simulate $N = 10,000$ observations $Y_1, \ldots, Y_N \sim \mathcal{B}e(p)$. Then

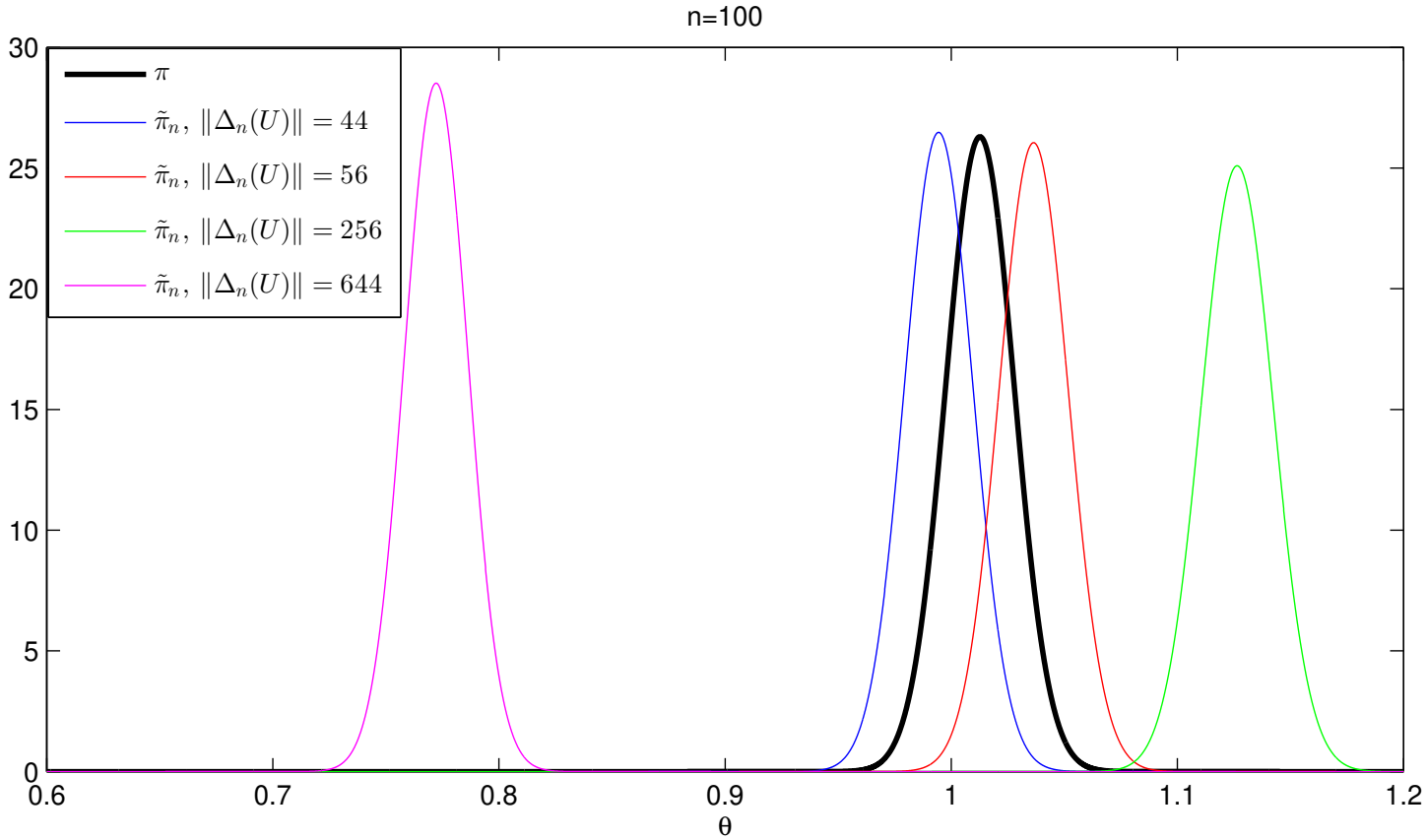
$$\pi(\theta \,|\, Y) \propto \pi(\theta)(1-p)^{\sum_{k=1}^{N} Y_k} p^{N - \sum_{k=1}^{N} Y_k}$$

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

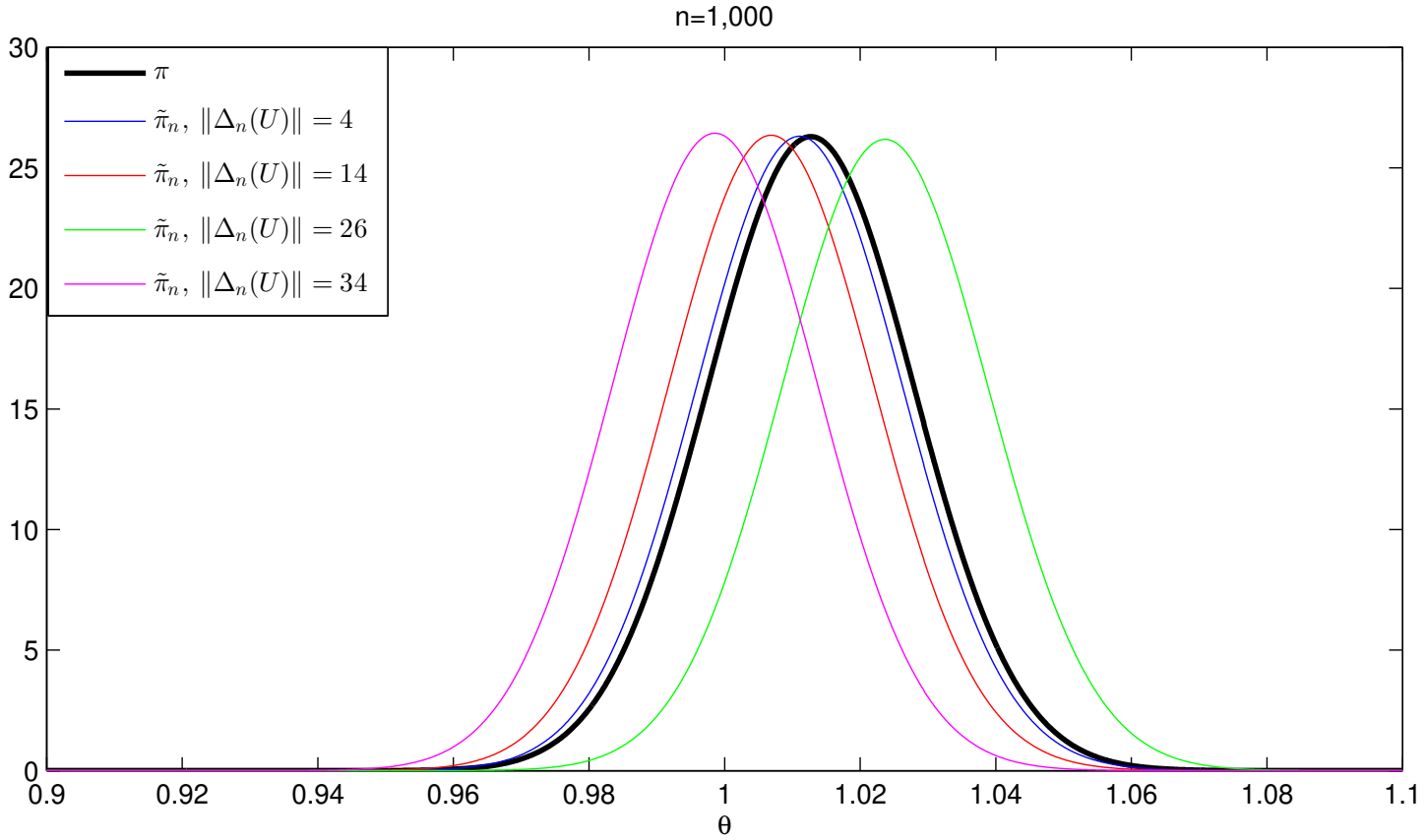
Here the sufficient statistics difference is

$$|\Delta_n(U)| = \left| \sum_{k=1}^{N} Y_k - \frac{N}{n} \sum_{k\in U} Y_k \right|.$$

Bernoulli model, $n = 100$



Bernoulli model, $n = 1000$



From sufficient statistics to summary statistics

We no longer assume that we are in an exponential model, nor i.i.d.

From sufficient statistics to summary statistics

We no longer assume that we are in an exponential model, nor i.i.d.

- Let S be a mapping of summary statistics.
- $\Delta_n(U) = S(Y_1, \ldots, Y_N) - \frac{N}{n} S(Y_U)$.

From sufficient statistics to summary statistics

We no longer assume that we are in an exponential model, nor i.i.d.

- Let S be a mapping of summary statistics.
- $\Delta_n(U) = S(Y_1, \ldots, Y_N) - \frac{N}{n} S(Y_U)$.

For each $U \in \mathcal{U}_n$, a *weight* $\nu_{n,\epsilon}(U)$ is assigned to the subset of data Y_U

$$\nu_{n,\epsilon}(U) \propto \exp\left\{-\epsilon \|\Delta_n(U)\|^2\right\} .$$

- $\epsilon \to 0$: all the subsets have the same weight.
- $\epsilon \to \infty$: the mass is concentrated on the most representative subset.

ISS-MCMC

Based on the analysis of exponential models, we propose ISS-MCMC (for *Informed Sub-Sampling*). It produces a chain $\{\theta_k, U_k\}_k$. A single iteration of this algorithm is as follows :

ISS-MCMC

Based on the analysis of exponential models, we propose ISS-MCMC (for *Informed Sub-Sampling*). It produces a chain $\{\theta_k, U_k\}_k$. A single iteration of this algorithm is as follows :

1. Update the subset :
   1. propose $U' \sim R(U_k, \cdot)$
   2. set U_{k+1} with probability
      $1 \wedge \exp\left\{\epsilon\left(\|\Delta_n(U_k)\| - \|\Delta_n(U')\|\right)\right\}$

ISS-MCMC

Based on the analysis of exponential models, we propose ISS-MCMC (for *Informed Sub-Sampling*). It produces a chain $\{\theta_k, U_k\}_k$. A single iteration of this algorithm is as follows :

1. Update the subset :
   1. propose $U' \sim R(U_k, \cdot)$
   2. set U_{k+1} with probability
      $1 \wedge \exp\left\{\epsilon\left(\|\Delta_n(U_k)\| - \|\Delta_n(U')\|\right)\right\}$
2. Update the parameter :
   1. propose $\theta' \sim Q(\theta_k, \cdot)$
   2. set $\theta_{k+1} = \theta'$ with probability

$$1 \wedge \frac{f(\theta' \mid Y_{U_{k+1}})\pi(\theta')Q(\theta', \theta_k)}{f(\theta_k \mid Y_{U_{k+1}})\pi(\theta_k)Q(\theta_k, \theta')}.$$

Assumptions for the analysis of ISS-MCMC

By construction, ISS-MCMC samples a Markov chain on an extended state space $\{(\theta_i, U_i), i \in \mathbb{N}\}$. Here we study the distribution of the marginal chain (θ_i).

Assumptions for the analysis of ISS-MCMC

By construction, ISS-MCMC samples a Markov chain on an extended state space $\{(\theta_i, U_i), i \in \mathbb{N}\}$. Here we study the distribution of the marginal chain (θ_i).

Some assumptions :

Assumptions for the analysis of ISS-MCMC

By construction, ISS-MCMC samples a Markov chain on an extended state space $\{(\theta_i, U_i), i \in \mathbb{N}\}$. Here we study the distribution of the marginal chain (θ_i).

Some assumptions :

1. We assume that the MH transition kernel K targetting π is either uniformly ergodic or geometrically ergodic.

Assumptions for the analysis of ISS-MCMC

By construction, ISS-MCMC samples a Markov chain on an extended state space $\{(\theta_i, U_i), i \in \mathbb{N}\}$. Here we study the distribution of the marginal chain (θ_i).

Some assumptions :

1. We assume that the MH transition kernel K targetting π is either uniformly ergodic or geometrically ergodic.
2. The subsets $U_1, U_2, \ldots$ are independent and identically distributed under $\nu_{n,\epsilon}$.

Assumptions for the analysis of ISS-MCMC

By construction, ISS-MCMC samples a Markov chain on an extended state space $\{(\theta_i, U_i), i \in \mathbb{N}\}$. Here we study the distribution of the marginal chain (θ_i).

Some assumptions :

1. We assume that the MH transition kernel K targetting π is either uniformly ergodic or geometrically ergodic.
2. The subsets $U_1, U_2, \ldots$ are independent and identically distributed under $\nu_{n,\epsilon}$.
3. There exists a constant $\gamma_n < \infty$, such that for all $(\theta, U) \in \Theta \times \mathcal{U}_n$

$$|\log f(Y \mid \theta) - (N/n) \log f(Y_U \mid \theta)| \leq \gamma_n N \|S(Y) - S(Y_U)\| .$$

Theoretical analysis of ISS-MCMC

Theorem 1

When the chain produced by the (regular) Metropolis-Hastings algorithm is geometrically ergodic, then so is $(\tilde{\theta}_i)$.

Theoretical analysis of ISS-MCMC

Theorem 2

When the chain produced by MH is uniformly ergodic, then so is $(\tilde{\theta}_i)$, and its asymptotic distribution $\tilde{\pi}_{n,\epsilon}$ satisfies, for some $\kappa > 0$,

$$\|\tilde{\pi}_{n,\epsilon} - \pi(\cdot|Y)\|_{\mathrm{TV}} \leq \kappa \delta_{n,\varepsilon}$$

where $\delta_{n,\varepsilon}$ is provided by the Noisy-MCMC analysis :

$$\delta_{n,\varepsilon} = \sup_{\theta} \int \mathbb{E}_{U \sim \nu_{n,\epsilon}} \left| 1 - \frac{\Phi_U(\theta')}{\Phi_U(\theta)} \right| Q(\theta, \mathrm{d}\theta'),$$

$$\text{where } \Phi_U(\theta) = \frac{f(Y|\theta)}{f(Y_U|\theta)^{\frac{N}{n}}}.$$

Choice of summary statistics

1. Clearly, this approach was initially motivated by the ABC method

Choice of summary statistics

1. Clearly, this approach was initially motivated by the ABC method
2. Ideally the choice of S should guarantee that subsamples Y_U having a very small likelihood $f(Y_U \mid \theta)$ 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 $f(Y_U \mid \theta)$ to go to 0 faster than $\nu_{n,\epsilon}(U)$. This is ensured if Assumption 3 holds.

Choice of summary statistics

1. Clearly, this approach was initially motivated by the ABC method
2. Ideally the choice of S should guarantee that subsamples Y_U having a very small likelihood $f(Y_U \mid \theta)$ 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 $f(Y_U \mid \theta)$ to go to 0 faster than $\nu_{n,\epsilon}(U)$. This is ensured if Assumption 3 holds.
3. In situations where the maximum likelihood estimator $\theta^*(Y_{1:n})$ is easy and quick to evaluate numerically, we recommend setting $S(Y_{1:n}) = \theta^*(Y_{1:n})$. In the case of independent observations, setting the summary statistics as the maximum likelihood estimate is justified since it is possible to prove that this implies that Assumption 3 holds.

Implementing ISS-MCMC

- Subset size n : This choice is essentially related to the computational budget available to the user. In the following examples we have used $n \propto N^{1/2}$ which achieves a substantial computational gain at a price of a negligible asymptotic bias.

- Bandwidth parameter ϵ : In practice, ϵ needs to be very large and could potentially cause the algorithm to get stuck in 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%.

Example 1 : estimation of template shapes

Data are of handwritten digits (MNIST database)

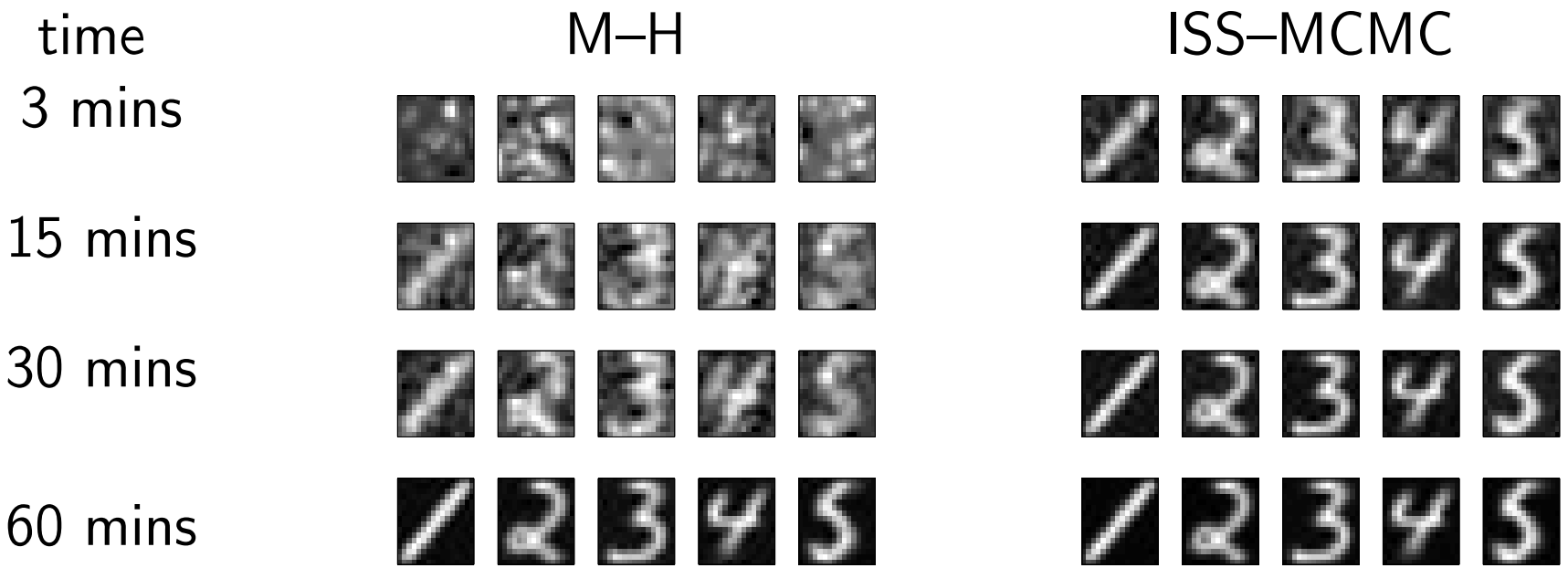
Figure – example of data

- The dataset contains $N = 10,000$ images of size 16x16
- Each image belongs to a class $I_k \in \{1, \ldots, 5\}$ assumed to be known
- The model can be written as :

$$I_k = i, \quad Y_k = \phi(\theta_i) + \sigma^2 \varepsilon_k \,, \qquad \varepsilon_k \sim \mathcal{N}(0, 1) \,.$$

Example 1 : estimation of template shapes

- $n = 100$ digits, $\epsilon = 10^5$ and $S(U) = (S_1(U), \ldots, S_5(U))$ with $S_i(U) = \sum_{k \in U} I_k$
- $\tau_{\text{MH}} = 41.2$ secs and $\tau_{\text{ISS-MCMC}} = 0.7$ secs ($60 \times$ faster)

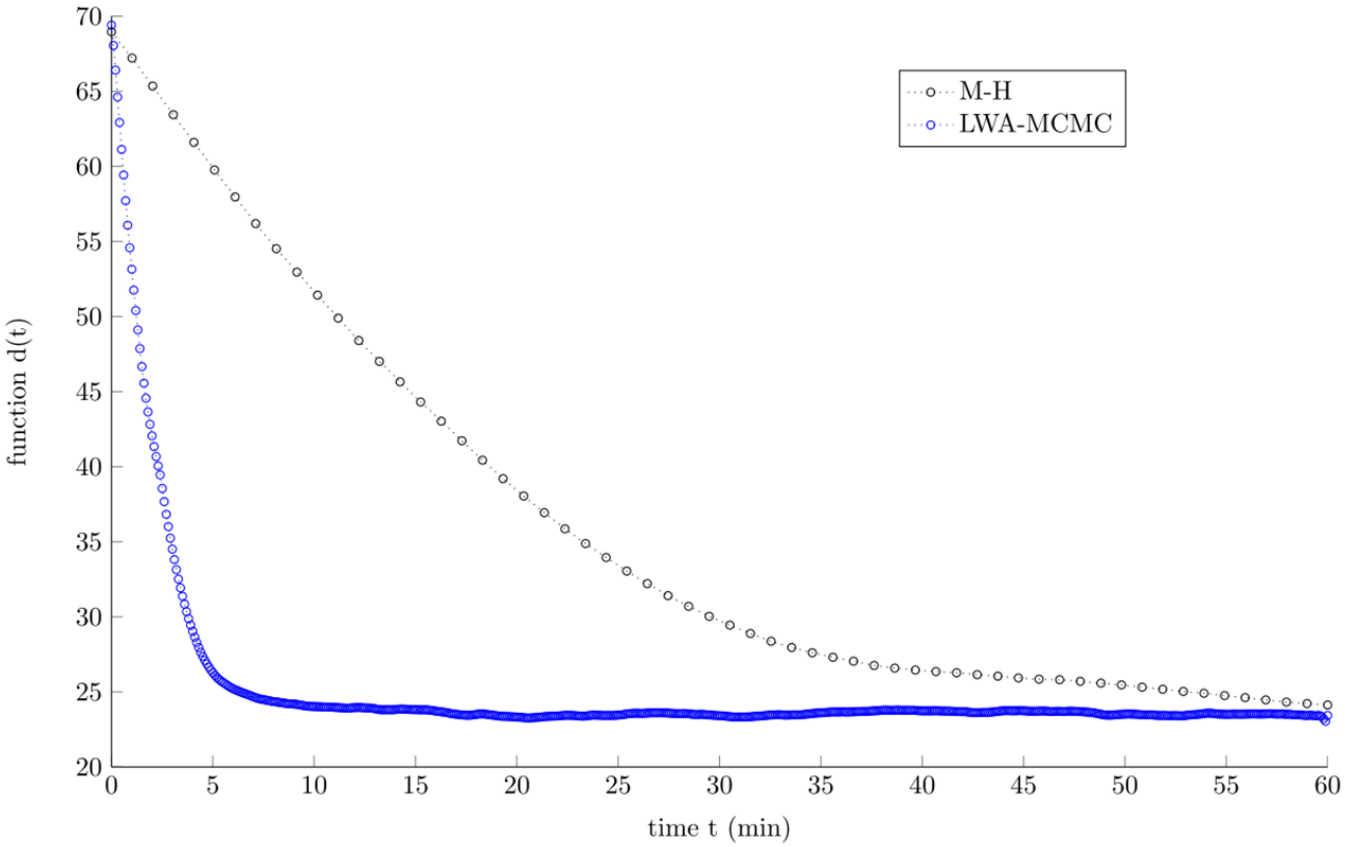


Example 1 : estimation of template shapes

Consider the metric $d(t) = \sum_{i=1}^{5} \left\| \theta_i^* - \frac{1}{L(t)} \sum_{\ell=1}^{L(t)} \theta_{i,\ell} \right\|$,

where :

- $L(t)$ is the number of iterations completed at time t
- θ_i^* is the map of model i (estimated from stochastic EM)



Example 1 : estimation of template shapes

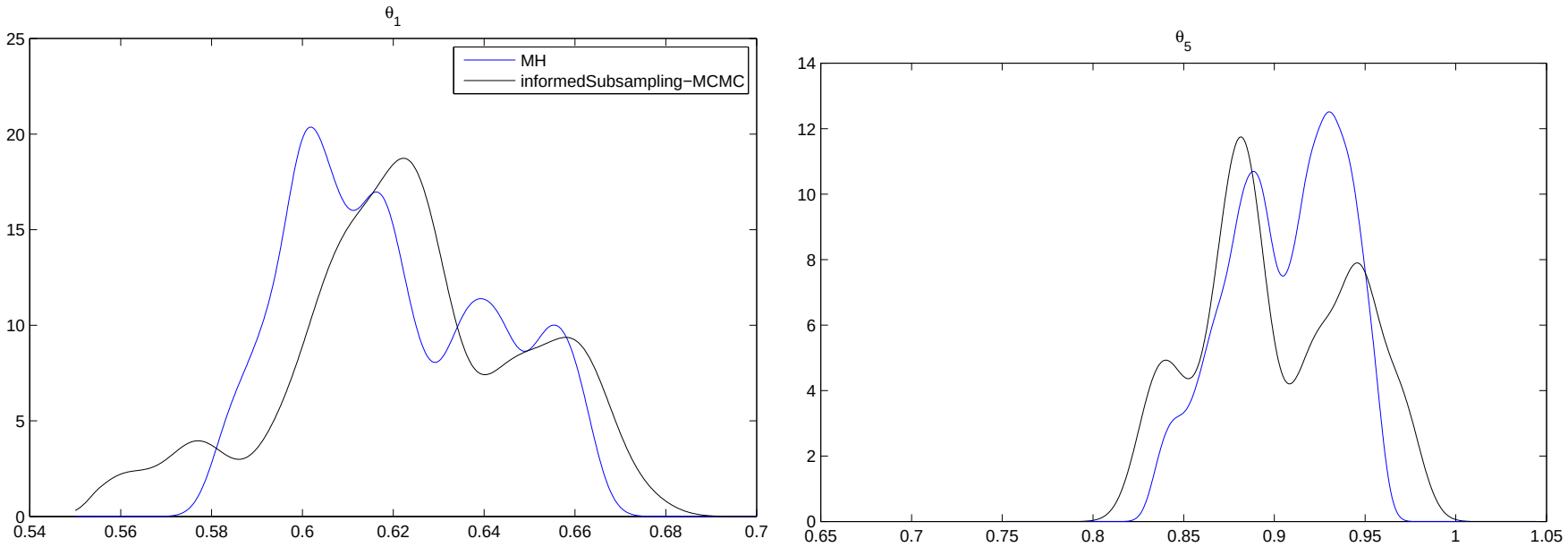


Figure – Comparing the true and approximate marginal distribution of one parameter of θ_1 (left) and one parameter of θ_5 (right).

Example 2 : Auto regressive model

An AR(2) model, parameterized by $\theta = (\theta_1 \theta_2, \theta_3)$

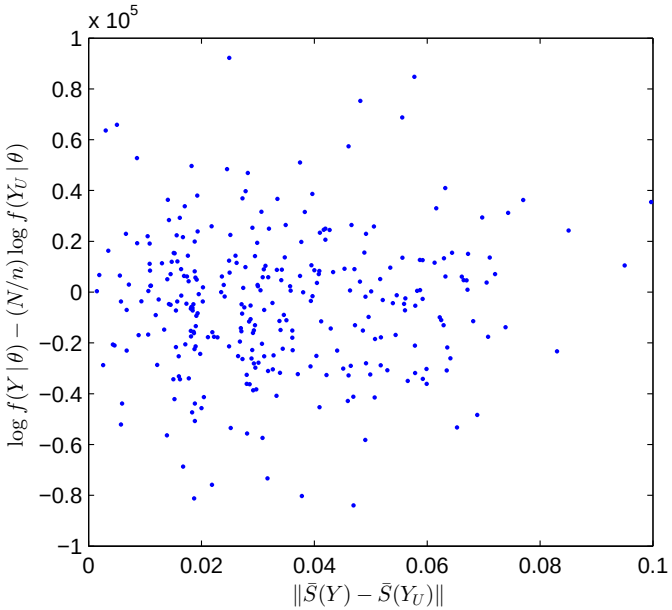
$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \zeta_t \,, \quad \zeta_t \sim \mathcal{N}(0, \theta_3^2) \,.$$

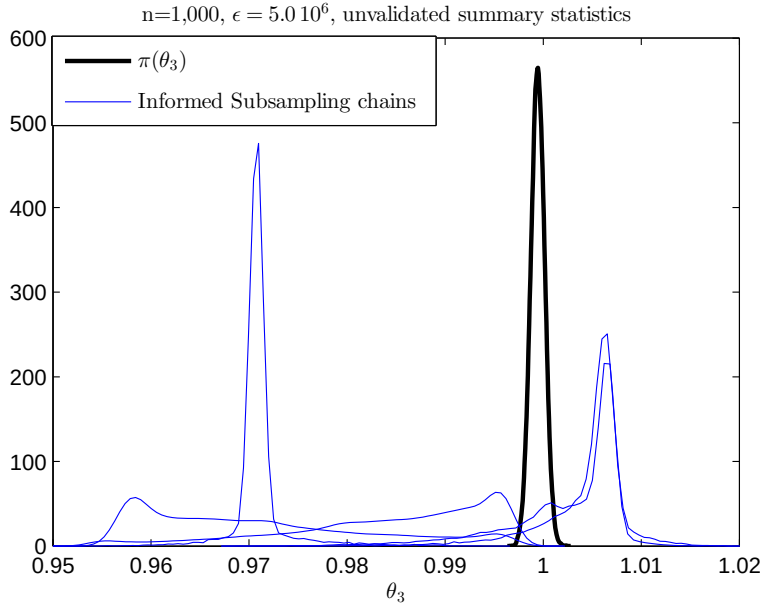
- An AR(2) time-series was simulated with $N = 10^6$ observations.
- Approximate inference with $n \in \{10^2, 10^3\}$.
- Summary statistics : auto-correlation function, or θ estimated by Yule-Walker ?
- Different ϵ were used.

Note : Sampling from $\pi(\,\cdot \,|\, y)$ via MH was a laborious task !

Example 2 : Auto regressive model

Case where S is defined as the estimated ACF (first 5 coefficients).

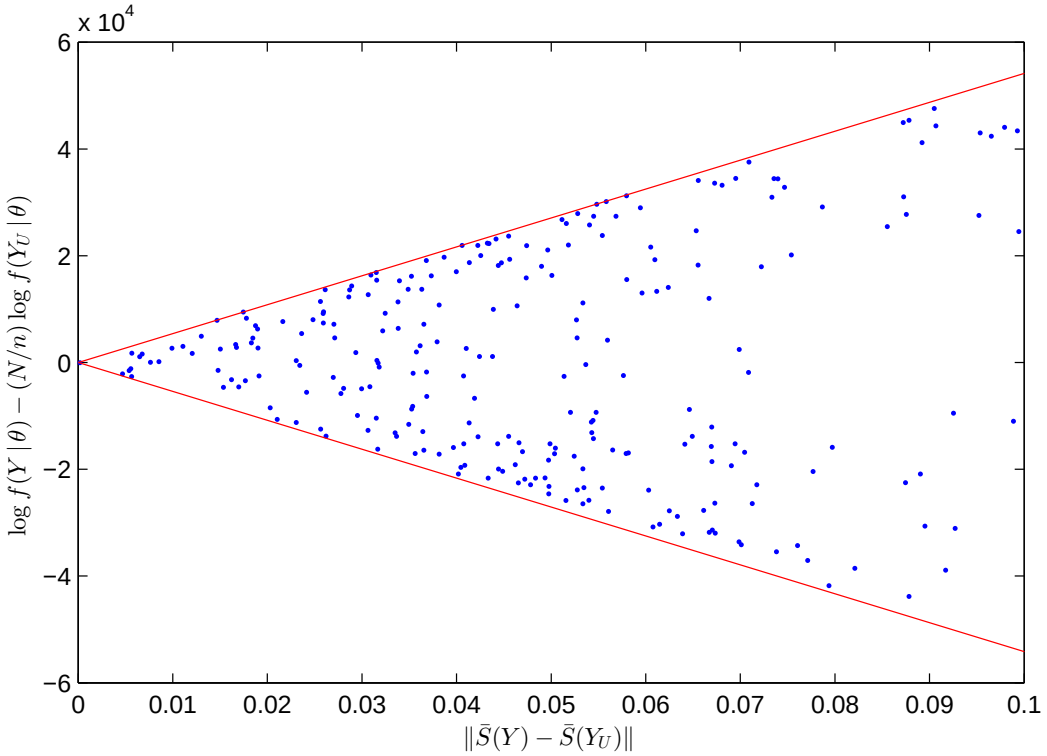




Here S is not useful, $\phi_U(\theta)$ is unstable yielding poor approximate posterior inference.

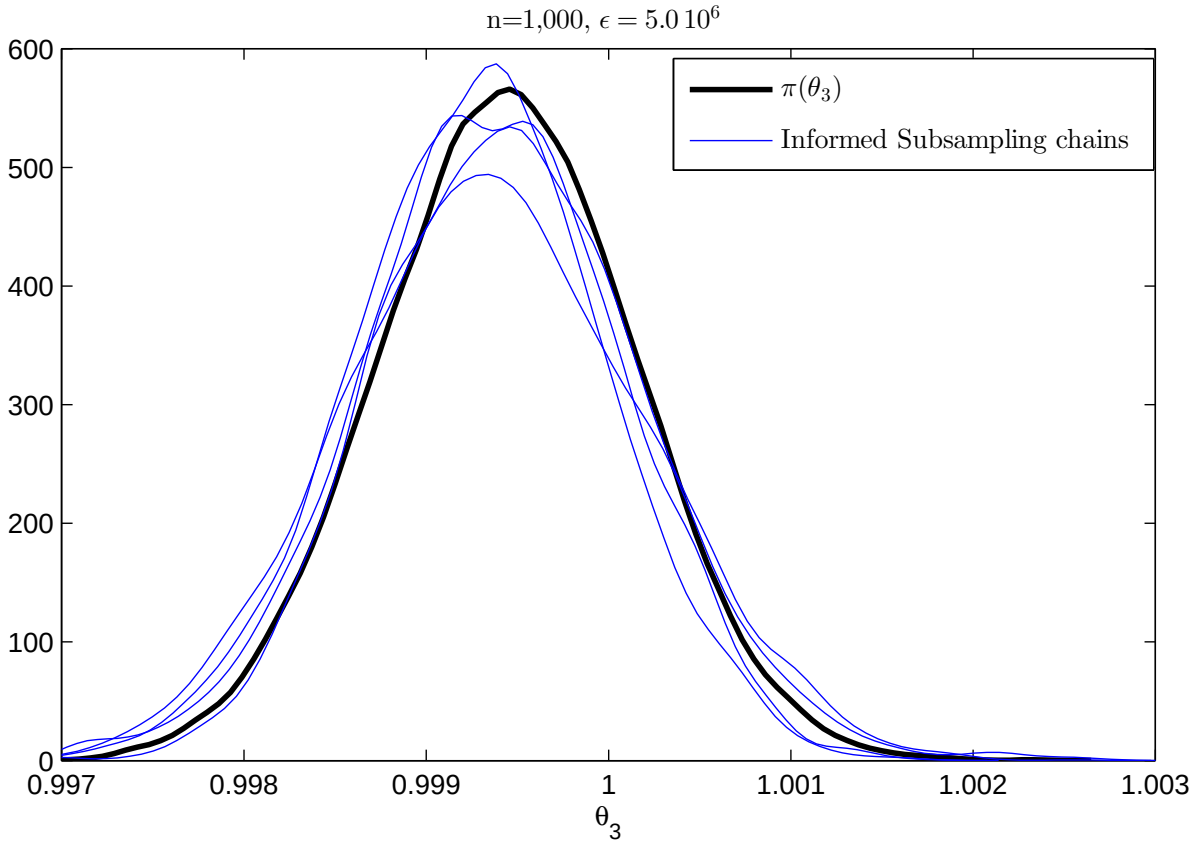
Example 2 : Auto regressive model

Second case is where S is defined as the Yule Walker coefs.



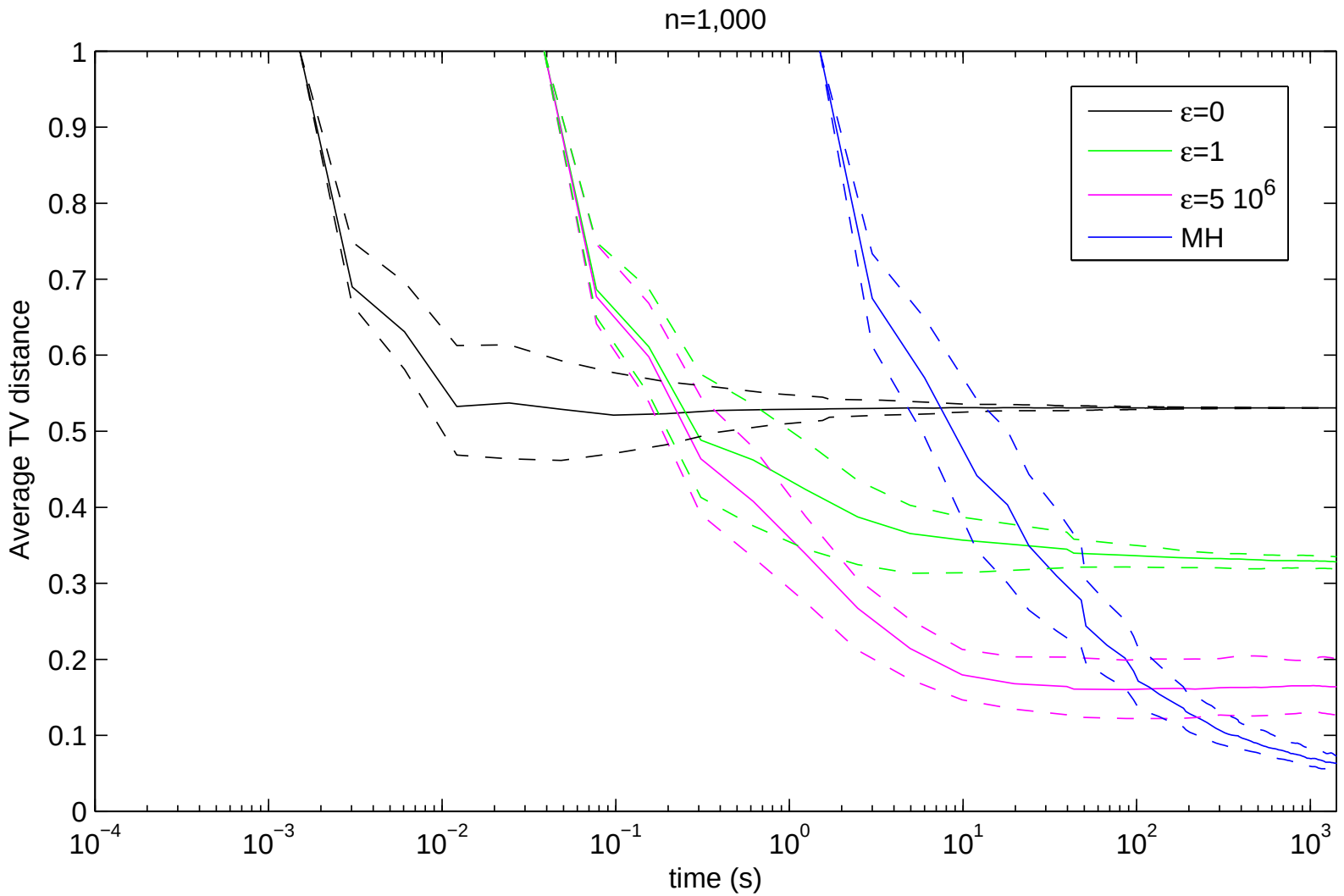
$\Rightarrow$ S is appropriate since the log ratio does not grow faster than linearly in $\|S(Y) - (N/n)S(Y_U)\|$.

Example 2 : Auto regressive model



Marginal distribution of θ_3 for 4 independent ISS-MCMC.

Example 2 : Auto regressive model



Conclusions

Our Informed Subsampling MCMC scheme :

- Allows one to control deterministically the MH transition complexity.
- Subsamples according to their fidelity to this full dataset, through summary statistics.
- Allows to control only asymptotically the distance between the chain distribution and the true posterior.

A recent paper

M. Quiroz, R. Kohn, M. Villani, M. Tran,. Speeding up MCMC by efficient data subsampling. *JASA*, to appear.

A recent paper

M. Quiroz, R. Kohn, M. Villani, M. Tran,. Speeding up MCMC by efficient data subsampling. *JASA*, to appear.

Pushed further the analysis of subsampling in Metropolis-Hastings with explicit rates. However, as far as I understand, the analysis only covers the $\epsilon = 0$ case.

A recent paper

M. Quiroz, R. Kohn, M. Villani, M. Tran,. Speeding up MCMC by efficient data subsampling. *JASA*, to appear.

Pushed further the analysis of subsampling in Metropolis-Hastings with explicit rates. However, as far as I understand, the analysis only covers the $\epsilon = 0$ case.

The website of the conference says that one of the authors should be here. I'd be happy to talk with you !

A word on Lanvegin-Monte-Carlo

Langevin-Monte Carlo approach can also be approximated by subsampling.

M. Welling, Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. *ICML*, 2011.

A word on Lanvegin-Monte-Carlo

Langevin-Monte Carlo approach can also be approximated by subsampling.

M. Welling, Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. *ICML*, 2011.

However, one has now two levels of approximation : discretization + subsampling. See the nice analysis

A. Dalalyan. Further and stronger analogy between sampling and optimization : Langevin Monte Carlo and gradient descent. *COLT*, 2017.

(requires strong assumptions on the distribution though).

A word on variational approximations

When one wants to get **really** faster, replacing sampling by optimization (Variational Bayes or EP) becomes unavoidable. Stochastic optimization then allows subsampling :


T. Broderick, N. Boyd, A. Wibisono, A. Wilson, M. Jordan. Streaming variational Bayes. *NIPS*, 2013.


M. Khan, D. Nielsen. Fast yet Simple Natural-Gradient Descent for Variational Inference in Complex Models. *ISITA*, 2018.

A word on variational approximations

When one wants to get **really** faster, replacing sampling by optimization (Variational Bayes or EP) becomes unavoidable. Stochastic optimization then allows subsampling :


T. Broderick, N. Boyd, A. Wibisono, A. Wilson, M. Jordan. Streaming variational Bayes. *NIPS*, 2013.


M. Khan, D. Nielsen. Fast yet Simple Natural-Gradient Descent for Variational Inference in Complex Models. *ISITA*, 2018.

The approximate posteriors were recently proven to be consistent (despite a possible deformation)


P. Alquier, J. Ridgway, N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 2016.


P. Alquier, J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *preprint arXiv*, 2017.


B.-E. Chérief-Abdellatif, P. Alquier. Consistency of Variational Bayes Inference for Estimation and Model Selection in Mixtures *Electronic Journal of Statistics*, 2018.


B.-E. Chérief-Abdellatif. Consistency of ELBO maximization for model selection. *AABI*, 2018.

Thank you !