



POLITECNICO
MILANO 1863

Coupled Markov chains with applications to Approximate Bayesian Computation for model based clustering

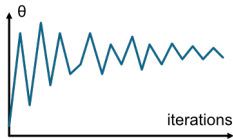
E. Bertoni, M. Caldarini, F. Di Filippo, G. Gabrielli, E. Musiani

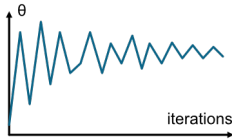
11 november 2021

Introduction

A complex problem

1/17



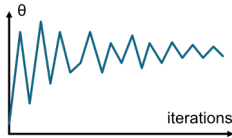


likelihood

Introduction

A complex problem

1/17

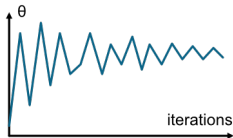


likely intractable

Introduction

A complex problem

1/17

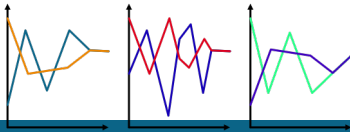


likelihood

intractable



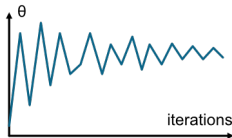
Unbiased Markov chain
Monte Carlo methods with
couplings



Introduction

A complex problem

1/17



likelihood

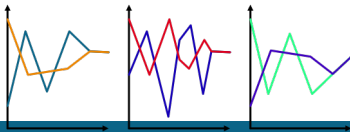
intractable




Unbiased Markov chain
Monte Carlo methods with
couplings



Approximate Bayesian
Computation





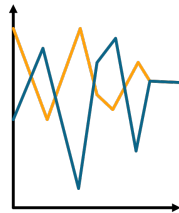
Unbiased Markov chain Monte Carlo methods with couplings

Faster MCMC \implies Parallelization

Faster MCMC \implies Parallelization \iff **Unbiased estimator**

Faster MCMC \implies Parallelization \iff **Unbiased estimator**

Exact estimations algorithms
using **coupling of Markov
chain.**



The goal is to estimate

$$\mathbb{E}_{\pi}[h(X)] = \int h(x)\pi(dx).$$

The estimator we are going to construct is based on a coupled pair of Markov chains, $(X_t)_{t \geq 0}$ and $(Y_t)_{t \geq 0}$, which marginally start from π_0 and evolve accordingly to P .

We consider some assumptions:

① as $t \rightarrow \infty$,

$$\mathbb{E}[h(X_t)] \rightarrow \mathbb{E}_\pi[h(X)];$$

and there exists $\eta > 0$ and $D < \infty$ such that $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$ for all $t \geq 0$;

We consider some assumptions:

- ① as $t \rightarrow \infty$,

$$\mathbb{E}[h(X_t)] \rightarrow \mathbb{E}_\pi[h(X)];$$

and there exists $\eta > 0$ and $D < \infty$ such that $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$ for all $t \geq 0$;

- ② the chains are such that the meeting time

$$\tau = \inf\{t \geq 1 : X_t = Y_{t-1}\}$$

satisfies $\mathbb{P}(\tau > t) \leq C\delta^t$ for all $t \geq 0$, for some constants $C < \infty$ and $\delta \in (0, 1)$;

We consider some assumptions:

- ① as $t \rightarrow \infty$,

$$\mathbb{E}[h(X_t)] \rightarrow \mathbb{E}_\pi[h(X)];$$

and there exists $\eta > 0$ and $D < \infty$ such that $\mathbb{E}[|h(X_t)|^{2+\eta}] \leq D$ for all $t \geq 0$;

- ② the chains are such that the meeting time

$$\tau = \inf\{t \geq 1 : X_t = Y_{t-1}\}$$

satisfies $\mathbb{P}(\tau > t) \leq C\delta^t$ for all $t \geq 0$, for some constants $C < \infty$ and $\delta \in (0, 1)$;

- ③ the chains stay together after meeting:

$$X_t = Y_{t-1} \text{ for all } t \geq \tau.$$

Thanks to the previous assumptions we can prove that:

$$\mathbb{E}_{\pi}[h(X)] = \mathbb{E}[h(X_k) + \sum_{t=k+1}^{\tau-1} \{h(X_t) - h(Y_{t-1})\}];$$

Thanks to the previous assumptions we can prove that:

$$\mathbb{E}_{\pi}[h(X)] = \mathbb{E}[h(X_k) + \sum_{t=k+1}^{\tau-1} \{h(X_t) - h(Y_{t-1})\}];$$

and we define the Rhee–Glynn estimator as:

$$H_k(X, Y) = h(X_k) + \sum_{t=k+1}^{\tau-1} \{h(X_t) - h(Y_{t-1})\}$$

which is **unbiased** by construction.

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{l=k}^m h(X_l) + \sum_{l=k+1}^{\tau-1} \min(1, \frac{l - k}{m - k + 1}) \{h(X_l) - h(Y_{l-1})\}$$

$$H_{k:m}(X, Y) = \underbrace{\frac{1}{m-k+1} \sum_{l=k}^m h(X_l)}_{MCMC_{k:m}} + \sum_{l=k+1}^{\tau-1} \min(1, \frac{l-k}{m-k+1}) \{h(X_l) - h(Y_{l-1})\}$$

- $MCMC_{k:m}$ is the standard MCMC average;

$$H_{k:m}(X, Y) = \underbrace{\frac{1}{m-k+1} \sum_{l=k}^m h(X_l)}_{MCMC_{k:m}} + \underbrace{\sum_{l=k+1}^{\tau-1} \min(1, \frac{l-k}{m-k+1}) \{h(X_l) - h(Y_{l-1})\}}_{BC_{k:m}}$$

- $MCMC_{k:m}$ is the standard MCMC average;
- $BC_{k:m}$ is the bias correction;

Time-averaged estimator II

- ① draw X_0 and Y_0 from an initial distribution π_0 and draw $X_1 \sim P(X_0, \cdot)$;
- ② set $t = 1$: while $t < \max\{m, \tau\}$ and:
 - a draw $(X_{t+1}, Y_t) \sim \bar{P}\{(X_t, Y_{t-1}), \cdot\}$;
 - b set $t \leftarrow t + 1$;
- ③ compute the time-averaged estimator:

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{l=k}^m h(X_l) + \sum_{l=k+1}^{\tau-1} \min(1, \frac{l - k}{m - k + 1}) \{h(X_l) - h(Y_{l-1})\}.$$

Time-averaged estimator II

- ① draw X_0 and Y_0 from an initial distribution π_0 and draw $X_1 \sim P(X_0, \cdot)$;
- ② set $t = 1$: while $t < \max\{m, \tau\}$ and:
 - a draw $(X_{t+1}, Y_t) \sim \bar{P}\{(X_t, Y_{t-1}), \cdot\}$; \bar{P} must be evaluated before!
 - b set $t \leftarrow t + 1$;
- ③ compute the time-averaged estimator:

$$H_{k:m}(X, Y) = \frac{1}{m - k + 1} \sum_{l=k}^m h(X_l) + \sum_{l=k+1}^{\tau-1} \min(1, \frac{l - k}{m - k + 1}) \{h(X_l) - h(Y_{l-1})\}.$$

Metropolis–Hasting algorithm allow us to calculate the coupled kernel $\bar{P}\{(X_t, Y_{t-1}), \cdot\}$:

- ① sample $(X^*, Y^*)|(X_t, Y_{t-1})$ from a maximal coupling of $q(X_t, \cdot)$ and $q(Y_{t-1}, \cdot)$;
- ② sample $U \sim \mathcal{U}([0, 1])$;

③ if

$$U \leq \min \left\{ 1, \frac{\pi(X^*)q(X^*, X_t)}{\pi(X_t)q(X_t, X^*)} \right\}$$

then $X_{t+1} = X^*$; otherwise $X_t = X_{t-1}$;

④ if

$$U \leq \min \left\{ 1, \frac{\pi(Y^*)q(Y^*, Y_t)}{\pi(Y_t)q(Y_t, Y^*)} \right\}$$

then $Y_{t+1} = Y^*$; otherwise $Y_t = Y_{t-1}$.

Metropolis–Hasting algorithm allow us to calculate the coupled kernel $\bar{P}\{(X_t, Y_{t-1}), \cdot\}$:

- 1 sample $(X^*, Y^*)|(X_t, Y_{t-1})$ from a maximal coupling of $q(X_t, \cdot)$ and $q(Y_{t-1}, \cdot)$;
- 2 sample $U \sim \mathcal{U}([0, 1])$;

- 3 if

$$U \leq \min \left\{ 1, \frac{\pi(X^*)q(X^*, X_t)}{\pi(X_t)q(X_t, X^*)} \right\}$$

then $X_{t+1} = X^*$; otherwise $X_t = X_{t-1}$;

- 4 if

$$U \leq \min \left\{ 1, \frac{\pi(Y^*)q(Y^*, Y_t)}{\pi(Y_t)q(Y_t, Y^*)} \right\}$$

then $Y_{t+1} = Y^*$; otherwise $Y_t = Y_{t-1}$.



Approximate Bayesian Computation

Inputs:

- a target posterior density $\pi(\theta|y_{obs}) \propto p(y_{obs}|\theta)\pi(\theta)$, consisting of a prior distribution $\pi(\theta)$ and a procedure of generating data under the model $p(y_{obs}|\theta)$;
- a proposal density $g(\theta)$, with $g(\theta) > 0$ if $\pi(\theta|y_{obs}) > 0$;
- an integer $N > 0$.

Inputs:

- a target posterior density $\pi(\theta|y_{obs}) \propto p(y_{obs}|\theta)\pi(\theta)$, consisting of a prior distribution $\pi(\theta)$ and a procedure of generating data under the model $p(y_{obs}|\theta)$;
- a proposal density $g(\theta)$, with $g(\theta) > 0$ if $\pi(\theta|y_{obs}) > 0$;
- an integer $N > 0$.

Sampling for $i = 1, \dots, N$:

- ① generate $\theta^{(i)} \sim g(\theta)$ from sampling density g ;
- ② generate $y \sim p(y|\theta^{(i)})$ from the likelihood;
- ③ if $y = y_{obs}$, then accept $\theta^{(i)}$ with probability $\frac{\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$, where $K \geq \max_{\theta} \frac{\pi(\theta)}{g(\theta)}$; else go to ①.

Inputs:

- a target posterior density $\pi(\theta|y_{obs}) \propto p(y_{obs}|\theta)\pi(\theta)$, consisting of a prior distribution $\pi(\theta)$ and a procedure of generating data under the model $p(y_{obs}|\theta)$;
- a proposal density $g(\theta)$, with $g(\theta) > 0$ if $\pi(\theta|y_{obs}) > 0$;
- an integer $N > 0$.

Sampling for $i = 1, \dots, N$:

- ① generate $\theta^{(i)} \sim g(\theta)$ from sampling density g ;
- ② generate $y \sim p(y|\theta^{(i)})$ from the likelihood;
- ③ if $y = y_{obs}$, then accept $\theta^{(i)}$ with probability $\frac{\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$, where
 $K \geq \max_{\theta} \frac{\pi(\theta)}{g(\theta)}$; else go to ①.

Output:

- a set of parameter vectors $\theta^{(1)}, \dots, \theta^{(N)}$ which are samples from $\pi(\theta|y_{obs})$.

Is this an efficient method for complex analysis?

Is this an efficient method for complex analysis?

③ If $\|y - y_{obs}\| \leq h$, then accept $\theta^{(i)}$ with probability $\frac{\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$, where $K \geq \max_{\theta} \frac{\pi(\theta)}{g(\theta)}$; else go to ①.

$$\pi(\theta, y|y_{obs}) \propto \mathbb{1}(\|y - y_{obs}\| \leq h) p(y|\theta) \pi(\theta)$$

$$\Downarrow$$

$$\pi_{ABC}(\theta, y|y_{obs}) \propto K_h(u) p(y|\theta) \pi(\theta)$$

$$\pi(\theta, y|y_{obs}) \propto \mathbb{I}(\|y - y_{obs}\| \leq h) p(y|\theta) \pi(\theta)$$

$$\Downarrow$$

$$\pi_{ABC}(\theta, y|y_{obs}) \propto K_h(u) p(y|\theta) \pi(\theta)$$

Where we used a **standard smoothing kernel function**:

$$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right), \quad \text{with } u = \|y - y_{obs}\|$$

Is this feasible in practice?

Is this feasible in practice?

No, it's difficult to have $y \approx y_{obs}$: we should use a large h , obtaining a poor posterior approximation!

\Rightarrow use summary statistics $s = S(y)$

Critical decision: choice of summary statistics

Critical decision: choice of summary statistics

Dimension of summary statistics:

- large enough to contain as much as information about observed data as possible
- low enough to avoid curse of dimensionality of matching s and s_{obs}

\implies choose sufficient statistics, such that:

$$\pi(\theta|s_{obs}) \equiv \pi(\theta|y_{obs})$$

Distance measure: substantial impact on ABC algorithm efficiency

$$\|s - s_{obs}\| = (s - s_{obs})^\top \Sigma^{-1} (s - s_{obs})$$

Distance measure: substantial impact on ABC algorithm efficiency

$$\|s - s_{obs}\| = (s - s_{obs})^\top \Sigma^{-1} (s - s_{obs})$$

- $\Sigma = \text{identity matrix} \rightarrow \text{Euclidean distance}$
- $\Sigma = \text{diagonal matrix of non-zero weights} \rightarrow \text{Weighted Euclidean distance}$
- $\Sigma = \text{full covariance matrix of } s \rightarrow \text{Mahalanobis distance}$

Inputs:

- a target posterior density $\pi(\theta|y_{obs}) \propto p(y_{obs}|\theta)\pi(\theta)$, consisting of a prior distribution $\pi(\theta)$ and a procedure of generating data under the model $p(y_{obs}|\theta)$;
- a proposal density $g(\theta)$, with $g(\theta) > 0$ if $\pi(\theta|y_{obs}) > 0$;
- an integer $N > 0$;
- a kernel function $K_h(u)$ and a scale parameter $h > 0$;
- a low dimensional vector of summary statistics $s = S(y)$.

Sampling for $i = 1, \dots, N$:

- ① generate $\theta^{(i)} \sim g(\theta)$ from sampling density g ;
- ② generate $y \sim p(y|\theta^{(i)})$ from the likelihood;
- ③ compute summary statistic $s = S(y)$;
- ④ accept $\theta^{(i)}$ with probability $\frac{K_h(\|s - s_{obs}\|)\pi(\theta^{(i)})}{Kg(\theta^{(i)})}$, where
 $K \geq K_h(0) \max_{\theta} \frac{\pi(\theta)}{g(\theta)}$; else go to 1.

Output:

- a set of parameter vectors $\theta^{(1)}, \dots, \theta^{(N)} \sim \pi_{ABC}(\theta|S_{obs})$.



Conclusions

Our focus till now was to understand the fundamental concepts and collect the missing information.

The next step will be a **simple and separate implementation** of both solution to be tested on simulated data.

Further steps will consider the **integration** of both solution into a single implementation and the testing on more complex data.

Pierre Jacob, John O'Leary, and Yves Atchadé.

Unbiased markov chain monte carlo with couplings.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82, 08 2017.

Peter W. Glynn and Chang han Rhee.

Exact estimation for markov chain equilibrium expectations, 2014.

Jeffrey S. Rosenthal.

Faithful couplings of markov chains: Now equals forever.

Advances in Applied Mathematics, 18(3):372–381, 1997.

Dylan Cordaro.

Markov chain and coupling from the past.

2017.

Jinming Zhang.

Markov chains, mixing times and coupling methods with an application in social learning.

2020.

S. A. Sisson, Y. Fan, and M. A. Beaumont.

Overview of approximate bayesian computation, 2018.

Y. Fan and S. A. Sisson.

Abc samplers, 2018.

Dennis Prangle.

Summary statistics in approximate bayesian computation, 2015.