Noisy Markov chain Monte Carlo: efficient algorithms for approximate Bayesian inference

Florian Maire, University College Dublin

joint work with : Nial, Lampros, Alan, Aidan, Riccardo, Pierre & Julien

Context

- Monte Carlo methods
  Estimating integrals of functions $f : \mathsf{X} \subset \mathbb{R}^d \to \mathbb{R}$

$$I = \int f(x)\pi(\mathrm{d}x)$$

- Markov chain Monte Carlo
  When simulating *i.i.d.* draws from π is not feasible, simulate a Markov chain

$$X_{n+1} \sim K(X_n, \cdot)$$

- Metropolis-Hastings
  1. propose $\tilde{X} \sim Q(X_n, \cdot)$
  2. accept/reject *i.e* set $X_{n+1} = \tilde{X}$ with proba.

$$\alpha(X_n, X_{n+1}) = 1 \wedge \frac{\pi(\tilde{X})Q(\tilde{X}, X_n)}{\pi(X_n)Q(X_n, \tilde{X})}$$

  and set $X_{n+1} = X_n$ otherwise

Shortcomings of Metropolis-Hastings

Each transition

$$X_n \to X_{n+1}$$

requires calculating the target distribution's pdf *i.e* $\pi(\tilde{X})$

In Bayesian analysis, where

$$\pi \equiv \pi(\theta \,|\, y) \propto f(y \,|\, \theta) p(\theta)$$

problems arise when

1. the likelihood function is not tractable
   $\Rightarrow$ MH can simply be not used!
2. the likelihood function evaluation is prohibitively slow
   $\Rightarrow$ MH can be implemented but will potentially generate only a few Markov chain samples for a (very) long run time

Outline

Outline

Outlines

Noisy Metropolis-Hastings

- In M-H, conditionally on $(X_n, \tilde{X})$, the acceptance ratio

$$a(X_n, \tilde{X}_n) = \frac{\pi(\tilde{X})Q(\tilde{X}, X_n)}{\pi(X_n)Q(X_n, \tilde{X})} \tag{1}$$

  is deterministic
- noisy M-H: replace $a(X_n, \tilde{X})$ by an estimator $\hat{a}(X_n, \tilde{X})$
  $\Rightarrow$ conditionally on $(X_n, \tilde{X})$, $\hat{a}(X_n, \tilde{X}_n)$ is a random variable
- If π is intractable but can be estimated pointwise *i.e* for all X

$$\hat{\pi}(X) \approx \pi(X)$$

  e.g by Monte Carlo, then use

$$\hat{a}(X_n, \tilde{X}_n) = \frac{\hat{\pi}(\tilde{X})Q(\tilde{X}, X_n)}{\hat{\pi}(X_n)Q(X_n, \tilde{X})}$$ [1]

[1] the idea was initially suggested by O'Neill, Philip D., et al. "Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods." Journal of the Royal Statistical Society: Series C (Applied Statistics) 49.4 (2000): 517-542.

Consequences of using a *noisy* acceptance ratio

- Metropolis-Hastings is by construction π-reversible and thus π invariant *i.e*

$$\text{if } X_0 \sim \pi \text{ and } X_1 \sim K(X_0, \cdot) \Rightarrow X_1 \sim \pi$$

or $\int \pi(\mathrm{d}x_0) K(x_0, \mathrm{d}x_1) = \pi(\mathrm{d}x_1)$

- Replacing $a(x, y)$ by a random variable $\hat{a}(x, y)$ induces a noise and in general,

$$\text{if } X_0 \sim \pi \text{ and } X_1 \sim \hat{K}(X_0, \cdot) \Rightarrow X_1 \not\sim \pi$$

- Two notorious exceptions where replacing $a(x, y)$ by $\hat{a}(x, y)$ gives a π-invariant chain:
  - the exchange algorithm (Murray et al, 2006)
  - the pseudo-marginal algorithm (Roberts and Andrieu, 2009)

We assume hereafter that the noisy chains considered are not π-invariant

Outlines

Computationally efficient algorithms (1/2)

At first glimpse, noisy MCMC appears useless since even asymptotically

$$\|\mathbb{P}\{\hat{X}_n \in \cdot\} - \pi\| \not\to 0$$

However, at best one would hope that an *exact* Markov chain is uniformly ergodic

$$\|\mathbb{P}\{X_n \in \cdot\} - \pi\| \leq C\rho^n$$

i.e drawing from π asymptotically, only

Computationally efficient algorithms (2/2)

Now, suppose that

- we can have a noisy Markov chains that satisfies

$$\|\mathbb{P}\{\hat{X}_n \in \cdot\} - \pi\| \leq \zeta + C\rho^n$$

  where ζ is a constant that results from using $\hat{a}$ instead of a.
- CPU time to generate one draw from the noisy chain $\hat{\tau}$ (compared to an exact chain τ) satisfies

$$\hat{\tau} \ll \tau \qquad e.g \quad \tau = \kappa\hat{\tau} \quad \kappa \gg 1$$

Then for a given amount of CPU time:

$$\|\mathbb{P}\{\hat{X}_{\kappa n} \in \cdot\} - \pi\| \leq \zeta + C\rho^{\kappa n} \leq C\rho^n = \|\mathbb{P}\{\hat{X}_n \in \cdot\} - \pi\|$$

$\Rightarrow$ the error term ζ is balanced out by the faster decay of the exponential component.

Asymptotically efficient algorithm

- Suppose that the M-H ratio $a(x, y)$ is not tractable but another ratio $\tilde{a}(x, y)$ is and leaves the chain invariant *i.e*

$$(i)\ \tilde{X} \sim Q(x, \cdot) \qquad (ii)\ \text{accept } \tilde{X} \text{ w.p. } \tilde{a}(x, tx)$$

  is a π-invariant Markov chain

- Peskun (1973) showed that for a given Randow Walk proposal Q, the Metropolis-Hastings acceptance ratio a yields the smallest possible asymptotic variance *i.e* in the limit of large n

$$\text{var}_{MH,n}(f) := \frac{1}{n}\text{Var}\sum_{k=1}^{n} f(X_k) \leq \text{var}_{P,n}(f) := \frac{1}{n}\text{Var}\sum_{k=1}^{n} f(X_k)$$

- How is the asymptotic variance degraded when the $\tilde{a}$ is used instead of a?
  $\Rightarrow$ What if $\tilde{a} \gg a$?
- Noisy MCMC suggests that using an inexact acceptance ratio $\hat{a}$ that is approximates may allow to inherit the optimal asymptotic properties of MH.

Outline

Two main contexts of interest

$$\pi \equiv \pi(\theta \,|\, y) \propto f(y \,|\, \theta) p(\theta)$$

- Computational complexity of the likelihood

  $f(y \,|\, \theta)$ is very slow to evaluate

  Examples:
  - tall dataset contexts
  - missing data models
- Intractable likelihood

  $$f(y \,|\, \theta) = \frac{q(\theta, y)}{Z(\theta)}, \quad Z(\theta) := \int q(\theta, \mathrm{d}y) \text{ is intractable}$$

  Example:
  - Gibbs random fields

Outlines

Efficient MCMC for tall dataset (1/3)

Suppose $y = \{y_1, y_2, \ldots, y_n\}$ is a very large dataset and f has no sufficient statistics, *e.g* $\log f(y, \theta) = R(y, \theta)$ where R is non linear.

- Metropolis-Hastings acceptance ratio is

$$a(\theta, \theta') = \frac{p(\theta')Q(\theta', \theta)}{p(\theta)Q(\theta, \theta')} e^{\{R(y,\theta') - R(y,\theta)\}}$$

$$= * \frac{p(\theta')Q(\theta', \theta)}{p(\theta)Q(\theta, \theta')} e^{\sum_{k=1}^{N} \{R(y_k,\theta') - R(y_k,\theta)\}} \quad (2)$$

- In Informed Subsampling MCMC (Maire et al., 2017), we propose using

$$\hat{a}(\theta, \theta', U) = \frac{p(\theta')Q(\theta', \theta)}{p(\theta)Q(\theta, \theta')} e^{\{R(y_U,\theta') - R(y_U,\theta)\}}$$

$$= * \frac{p(\theta')Q(\theta', \theta)}{p(\theta)Q(\theta, \theta')} e^{\sum_{k \in U} \{R(y_k,\theta') - R(y_k,\theta)\}} \quad (3)$$

where $U \subset \{1, \ldots, N\}$ and $|U| \ll N$ such that y_U is a subsampling of y which is drawn such that y_U and y have similar summary statistics.

Efficient MCMC for tall dataset (2/3)

Consider the following logistic regression model with $N = 10^6$ observations

(i) simulate $X_{i,1}, \ldots, X_{i,N} \sim \mathcal{N}(0, (1/d)^2)$

(ii) simulate Y_i given θ and X_i as

$$Y_i = \begin{cases} 1 & \text{w.p. } 1/\left\{1 + e^{-\theta^T X}\right\} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

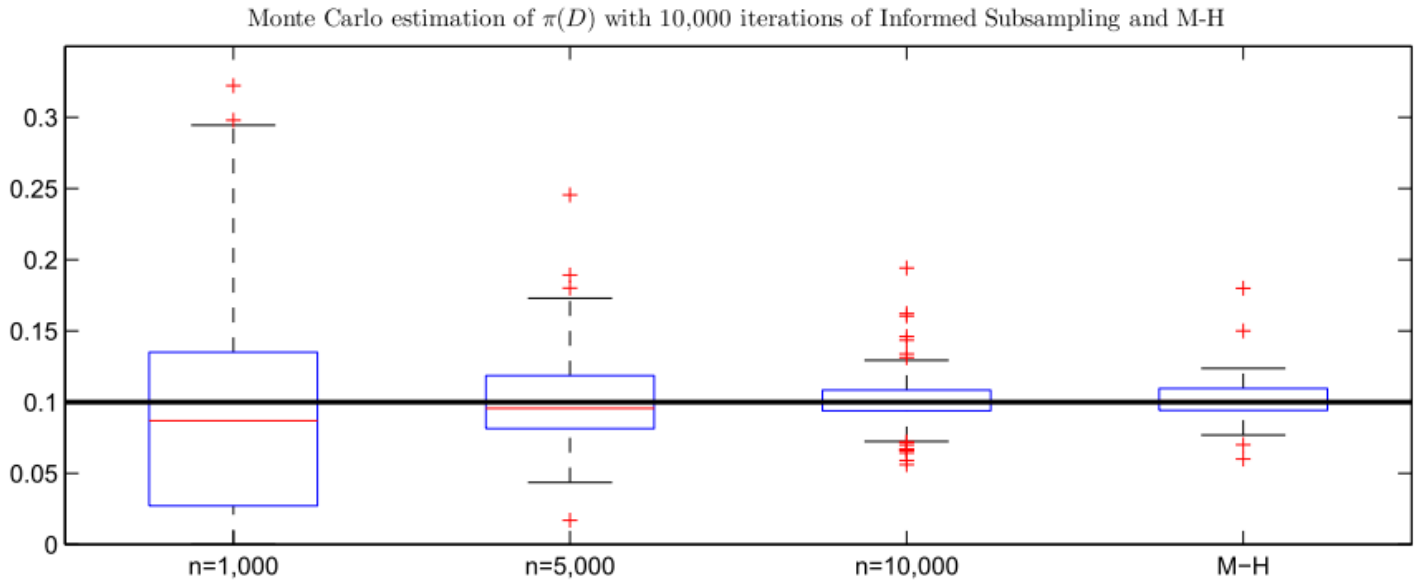
Goal: estimation of some probability

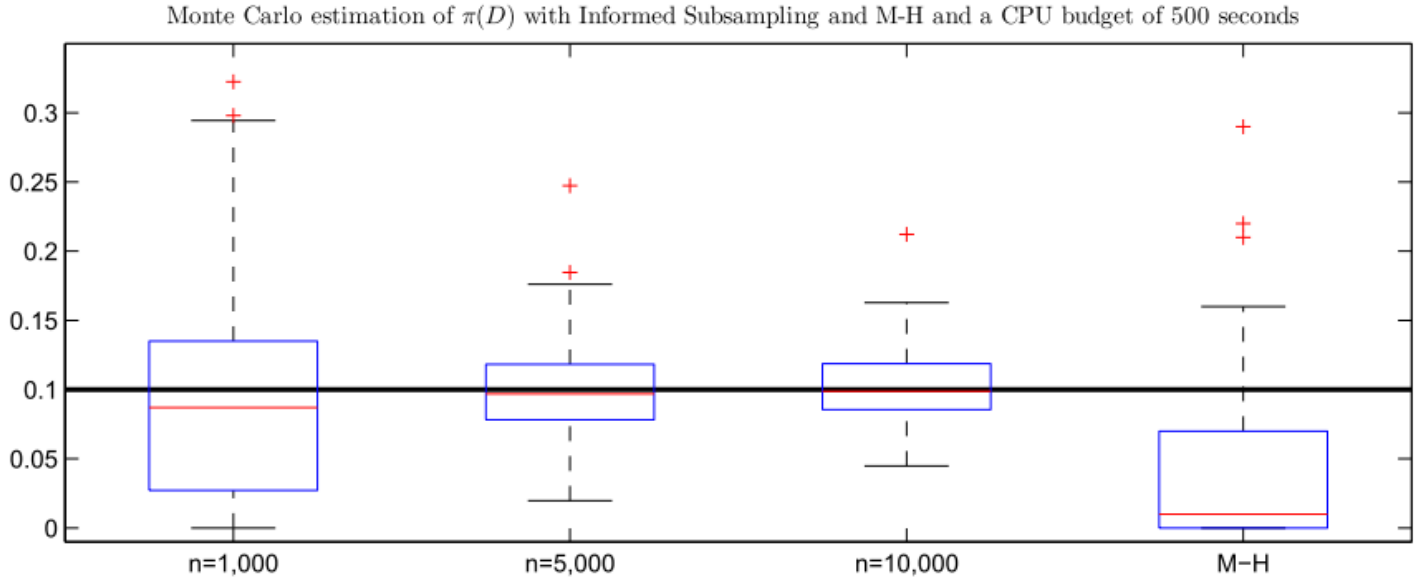
$$\pi\{D\} = \int \mathbb{1}_D(\theta)\pi(\mathrm{d}\theta \,|\, y)$$

algorithm	time/iter.(s)	iter. completed	RMSE	$\widehat{\mathrm{var}\{\pi(D)\}}$
M–H	10	50	0.1417	0.004
IIS-MCMC, $n = 1,000$	0.05	10,000	0.1016	0.0104
IIS-MCMC, $n = 5,000$	0.08	6,250	0.0351	0.0012
IIS-MCMC, $n = 10,000$	0.13	3,840	0.0267	0.0007

Figure: ISS-MCMC: our "fast" MCMC method with different subsampling size n

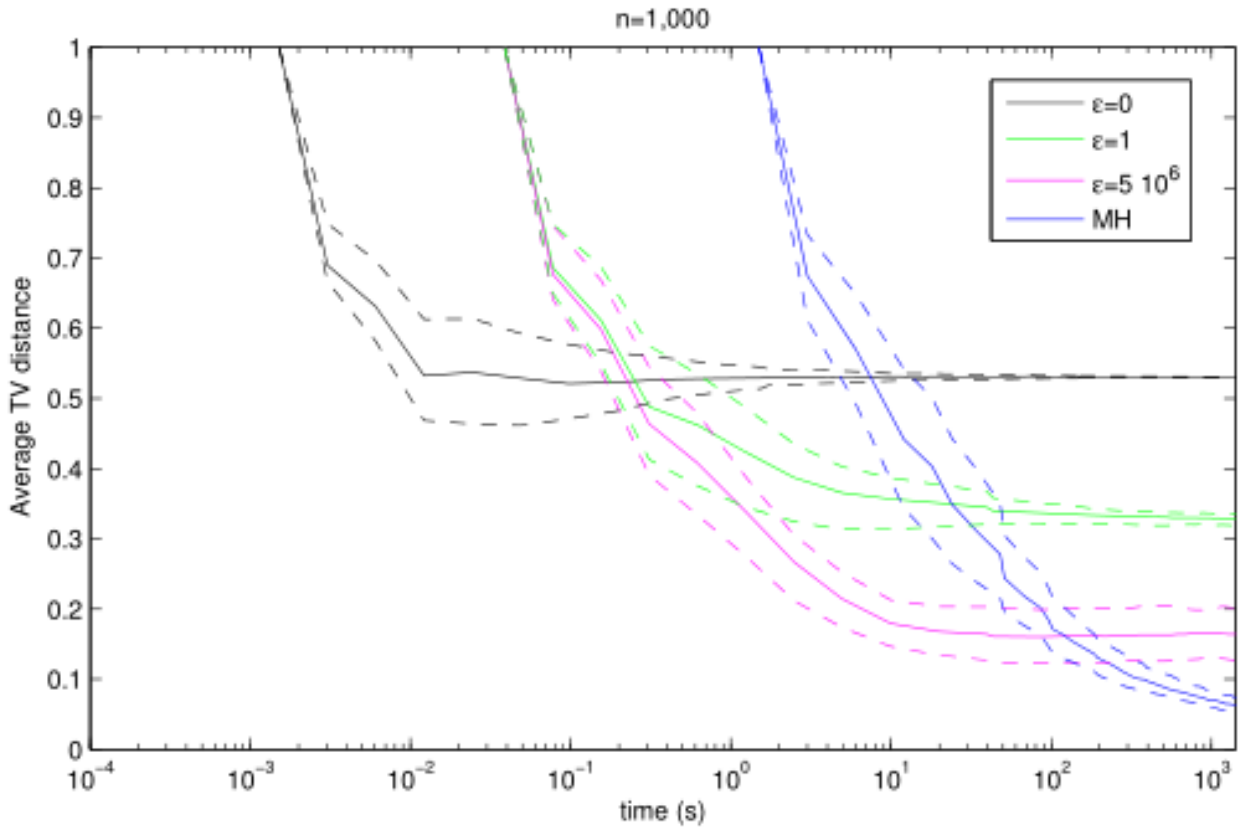
MCMC for tall dataset (3/3)





MCMC for tall dataset

- Subset of size $n = 1,000 \ll N = 10^6$
- Uniform subsampling ($\epsilon = 0$) and Informed Subsampling ($\epsilon \gg 1$)



Latent position model (Rastelli et al., 2017) (1/3)

- y is a network with N nodes and $\{y_{i,j}\}_{i \neq j}$ is the adjacency matrix
- a latent (unknown) position $z_i \in (-1, 1)^2$ is assigned to each node i
- the complete data likelihood writes

$$f(y \mid z, \theta) = \prod_{i<j} p(z_i, z_j; \theta)^{y_{i,j}} (1 - p(z_i, z_j; \theta))^{1-y_{i,j}} ,$$

$$p(z_i, z_j; \theta) = \frac{\exp\{\theta - \|z_i - z_j\|\}}{1 + \exp\{\theta - \|z_i - z_j\|\}} \quad (5)$$

  and CPU cost grows in $\mathcal{O}(N^2)$
- define a grid that spans $(-1, 1)^2$ in blocs $\mathbb{B} = \{\mathbb{b}_1, \ldots, \mathbb{b}_M\}$
- approximate the likelihood contribution of node 1

$$\prod_{j>1} p(z_1, z_j; \theta)^{y_{1,j}} (1 - p(z_1, z_j; \theta))^{1-y_{1,j}}$$

  by

$$\prod_{\mathbb{b}_k \in \mathbb{B}} p(z_1, \mathbb{b}_k; \theta)^{|\mathbb{b}_k(z)|} (1 - p(z_1, \mathbb{b}_k; \theta))^{1-|\mathbb{b}_k(z)|}$$

Latent position model (2/3)

- This approximation can be applied to the M-H acceptance ratio
- Approximate

$$a(\theta, z, \theta', z') = \frac{p(\theta', z')Q(\theta', z', \theta, z)}{p(\theta, z)Q(\theta, z, \theta', z')} \times \frac{f(y \mid \theta', z')}{f(y \mid \theta, z)}$$

$$\text{where} \quad f(y \mid z, \theta) = \prod_{i<j} p(z_i, z_j; \theta)^{y_{i,j}} (1 - p(z_i, z_j; \theta))^{1-y_{i,j}} = \mathcal{O}(N^2)$$

by

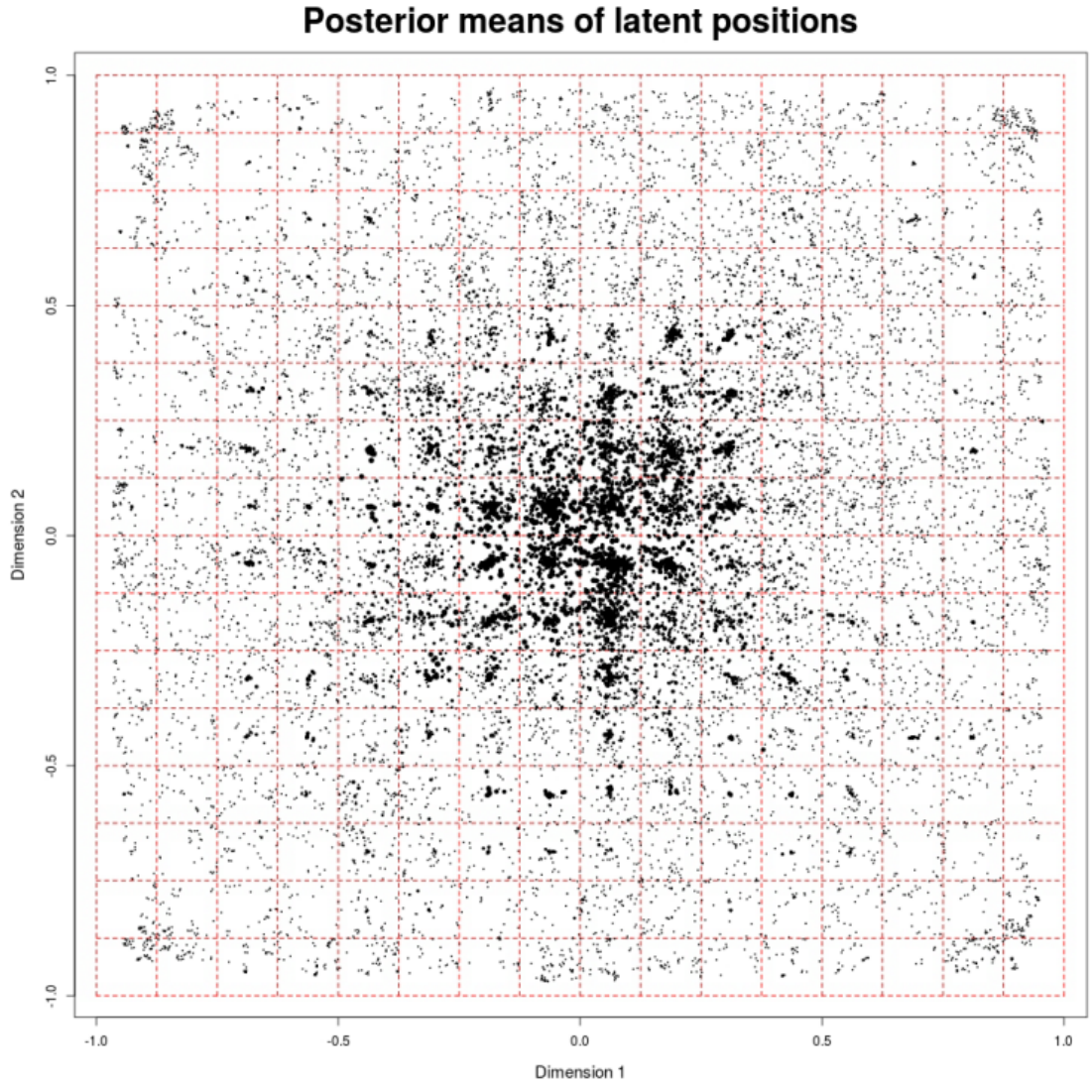
$$\hat{a}_{\mathbb{B}}(\theta, z, \theta', z') = \frac{p(\theta', z')Q(\theta', z', \theta, z)}{p(\theta, z)Q(\theta, z, \theta', z')} \times \frac{\hat{f}_{\mathbb{B}}(y \mid \theta', z')}{\hat{f}_{\mathbb{B}}(y \mid \theta, z)}$$

where

$$\hat{f}_{\mathbb{B}}(y \mid z, \theta) = \prod_{i=1}^{N} \prod_{\mathbb{b}_k \in \mathbb{B}} p(z_i, \mathbb{b}_k; \theta)^{|\mathbb{b}_k(z)|} (1 - p(z_i, \mathbb{b}_k; \theta))^{1-|\mathbb{b}_k(z)|} = \mathcal{O}(N)$$

Latent position model (3/3)

Co-authorship network (papers submitted on arXiv in the astro-physics category): $18,872$ authors



Outlines

The exchange algorithm

- When the likelihood $f(y \mid \theta) \propto q(y, \theta')$ has an intractable normalizing constant, an exact MH algorithm still exists:
- Replace

$$a(\theta, \theta') = \frac{q(y, \theta')p(\theta')Q(\theta', \theta)}{q(y, \theta)p(\theta)Q(\theta, \theta')} \times \frac{Z(\theta)}{Z(\theta')}$$

by

$$\hat{a}_{\mathsf{EX}}(\theta, \theta', y') = \frac{q(y, \theta')p(\theta')Q(\theta', \theta)}{q(y, \theta)p(\theta)Q(\theta, \theta')} \times \frac{q(y', \theta)}{q(y', \theta')} \qquad y' \sim f(\cdot \mid \theta')$$

- This algorithm is known as the exchange algorithm (Murray, 2006) and is *exact i.e*

$$\|\mathbb{P}_{\mathsf{EX}}\{X_n \in \cdot\} - \pi\| \to 0$$

Drawbacks of the exchange

- Empirical results have shown that the exchange suffers from large asymptotic variance (*i.e* low asymptotic efficiency)
- Peskun (1973) showed that for a given Randow Walk proposal Q, the Metropolis-Hastings acceptance ratio a yields the smallest possible asymptotic variance
  $\Rightarrow$ How is the asymptotic variance degraded when the exchange ratio is used?
- How to remedy to it?

The exchange inefficiency

- Roughly speaking, the Peskun ordering says that

  the larger the chance to move from current state

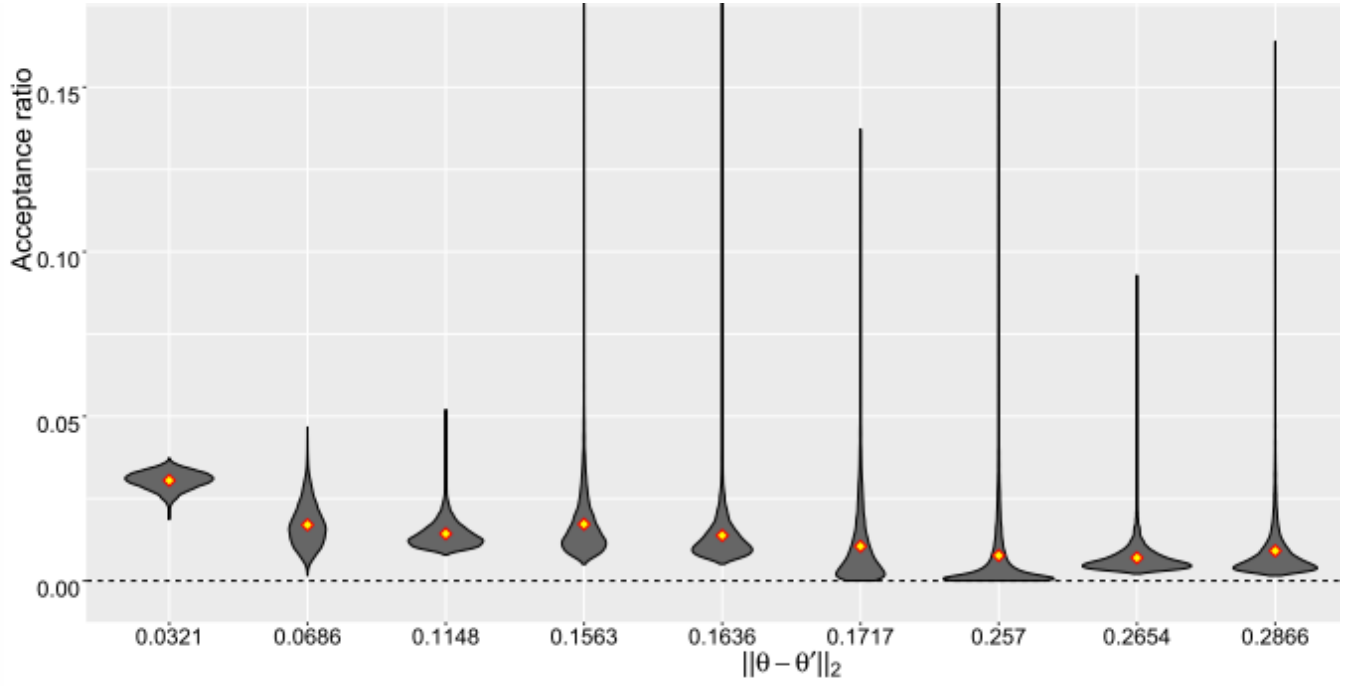
  $\Rightarrow$ the smallest the asymptotic variance

- We have established that

$$\mathbb{P}_{\theta,\theta'}\{a(\theta,\theta') \geq \hat{a}_{\mathsf{EX}}(\theta,\theta')\} \geq 1/2$$

  and that this probability increases with $\|\theta - \theta'\|$

- this is illustrated as follows for an exponential random graph model



Noisy Metropolis-Hastings for exponential random graphs

- finding an unbiased positive estimate of the likelihood normalizing constant is (very) challenging
  $\Rightarrow$ Pseudo-marginal algorithms cannot be used
- in some instances, finding an unbiased estimate of the **ratio** is possible, *e.g* for exponential models

$$q(y,\theta) = e^{g(\theta)^T S(y)} \Rightarrow \frac{Z(\theta)}{Z(\theta')} = \mathbb{E}_{f(\,\cdot\,|\,\theta')}\left\{\frac{e^{g(\theta)^T S(Y')}}{e^{g(\theta')^T S(Y')}}\right\}$$

- Replace

$$a(\theta,\theta') = \frac{q(y,\theta')p(\theta')Q(\theta',\theta)}{q(y,\theta)p(\theta)Q(\theta,\theta')} \times \frac{Z(\theta)}{Z(\theta')}$$

by

$$\hat{a}(\theta,\theta') = \frac{q(y,\theta')p(\theta')Q(\theta',\theta)}{q(y,\theta)p(\theta)Q(\theta,\theta')} \times \frac{1}{n}\sum_{i=1}^{n}\frac{q(y_i',\theta)}{q(y_i',\theta')}\,, \qquad y_1', y_2', \ldots \sim f(\,\cdot\,|\,\theta')$$

Illustration: Exponential Graph Models

- For this type of model, $q(y, \theta) = \exp\{\theta^T S(y)\}$ where S is a sufficient statistics vector that characterizes the model
- The Zachary Karate Club dataset with 2 statistics: # edges and # shared partnership (*i.e* # of connected nodes with exactly the same number of common neighbours)

		noisy MH			
type of algo	exchange	$n = 5$	$n = 100$	$n = 1,000$	$\star$
random walk proposal	1	1.3	1.6	1.8	1.9
independent proposal	1	1.4	2.3	3.0	3.6

Table: Effective sample size (relative to the exchange)

$\star$: noisy MH with a more sophisticated estimator of the normalizing constant.

Outline

Theoretical issues

The previous experiments have shown that using an approximation of the Metropolis-Hastings acceptance ratio

- allows to devise algorithms that have smaller computational complexity
- allows to devise algorithms that are asymptotically more efficient than existing ones

One central question remains to be addressed

- how can we trust the outputs of the those noisy chains *i.e*

$$\lim_{n\to\infty} \|\mathbb{P}\{\hat{X}_n \in \cdot\} - \pi\| \to \epsilon \ll 1 \ ?$$

- in particular, can they be used to form an estimator *i.e*

$$\lim_{n\to\infty} \left| \frac{1}{n} \sum_{k=1}^{n} f(\hat{X}_k) - I \right| \to \epsilon \ll 1 \ ?$$

Outlines

Uniformly ergodicity of K

Under uniform ergodicity assumption for the exact kernel K *i.e*

$$\sup_{x \in \mathsf{X}} \|\mathbb{P}\{X_n \in \cdot\} - \pi\| \leq C\rho^n \,,$$

Mitrophanov (2005) showed that for sufficiently large n

$$\sup_{x \in \mathsf{X}} \|K^n(x, \cdot) - \hat{K}^n(x, \cdot)\| \leq \left\{ \lambda + C\frac{\rho^\lambda - \rho^n}{1 - \rho} \right\} \sup_{x \in \mathsf{X}} \|K(x, \cdot) - \hat{K}(x, \cdot)\|$$

$(\lambda = \lceil \log_\rho C^{-1} \rceil)$

Alquier et al. 2016

Alquier et al. (2016) showed that for a M-H randomized acceptance ratio $\hat{a}_\eta(x, y)$

$$\sup_{x \in \mathsf{X}} \|K(x, \cdot) - \hat{K}(x, \cdot)\| \leq \sup_{x \in \mathsf{X}} \int Q(x, \mathrm{d}y) \mathbb{E}_\eta \left\{ |a(x, y) - \hat{a}_\eta(x, y)| \right\}$$

When $\hat{a}_{\eta,N}$ is a Monte Carlo estimate of a *i.e* $a(x, y) = \int a(x, y, \xi) f(\mathrm{d}\xi)$

$$\hat{a}_{\xi,N}(x, y) = \frac{1}{N} \sum_{k=1}^{N} a(x, y, \xi_k), \quad \xi_1, \xi_2, \ldots \sim_{i.i.d.} f$$

Then

$$\begin{aligned} \sup_{x \in \mathsf{X}} \|K^n(x, \cdot) - \hat{K}^n(x, \cdot)\| &\leq \left\{ \lambda + C \frac{\rho^\lambda - \rho^n}{1 - \rho} \right\} \Phi(Q, \pi) \mathrm{Var}\{\hat{a}_{\xi,N}(x, y)\}^{1/2} \\ &= \left\{ \lambda + C \frac{\rho^\lambda - \rho^n}{1 - \rho} \right\} \Phi(Q, \pi) \frac{1}{N} \mathrm{Var}\{a(x, y, \xi)\}^{1/2} \end{aligned}$$

Johndraw et al. (2017) (1/2)

The Authors work in the framework where

- K satisfies a Doeblin condition, *i.e* there exists $\kappa \in (0, 1)$ s.t.

$$\sup_{(x,y)\in \mathsf{X}^2} \|K(x, \cdot) - K(y, \cdot)\| \leq 1 - \kappa$$

- There exists $\gamma \in (0, 1)$ such that the approximate kernel satisfies

$$\sup_{x\in \mathsf{X}} \|K(x, \cdot) - \hat{K}(x, \cdot)\| \leq \gamma$$

If $\epsilon < \alpha/2$, $\hat{K}$ also satisfies a Doeblin condition and has an invariant measure $\hat{\pi}$ s.t.

$$\|\pi - \hat{\pi}\| \leq \gamma/\alpha\,.$$

Johndraw et al. (2017) (2/2)

The Authors work in the framework where

- K satisfies a Doeblin condition, *i.e* there exists $\kappa \in (0,1)$ s.t.

$$\sup_{(x,y)\in \mathsf{X}^2} \|K(x,\cdot) - K(y,\cdot)\| \leq 1-\kappa$$

- There exists $\gamma \in (0,1)$ such that the approximate kernel satisfies

$$\sup_{x\in\mathsf{X}} \|K(x,\cdot) - \hat{K}(x,\cdot)\| \leq \gamma$$

In this case, we have

$$\frac{1}{\|f\|_*^2}\mathbb{E}\left\{\int f(x)\mathrm{d}\pi(x) - \frac{1}{n}\sum_{k=0}^{n-1} f(\hat{X}_k)\right\}^2 \leq \frac{4\gamma^2}{\alpha^2} + \mathcal{O}(1/n)$$

for any function f satisfying $\|f\|_* := \inf_{\lambda\in\mathbb{R}} \sup_{x\in\mathsf{X}} |f(x)+\lambda| < \infty$.

Outlines

Results in Wasserstein distance

The Wasserstein distance between two measures (μ, ν) is defined as

$$W_d(\mu, \nu) = \inf_{\Gamma \in \mathcal{C}(\mu,\nu)} \iint d(x, y)\Gamma(\mathrm{d}\mu, \mathrm{d}\nu)$$

Assumption on K If there exists $C > 0$, $\rho \in (0, 1)$

$$\sup_{x \neq y} \frac{W_d(K^n(x, \cdot), K^n(y, \cdot))}{d(x, y)} \leq C\rho^n$$

Assumption on $\hat{K}$ There exists $V : \mathsf{X} \to (0, \infty)$, $\delta \in (0, 1)$ and $L > 0$

$$\hat{K}\hat{V}(x) \leq \delta\hat{V}(x) + L \qquad \forall\, x \in \mathsf{X}$$

then

$$W(p_n, \hat{p}_n) \leq C(1 - \rho^n)\frac{\gamma\kappa}{1 - \rho}$$

where

$$\gamma = \sup_{x \in \mathsf{X}} \frac{W_d(K(x, \cdot), \hat{K}(x, \cdot))}{\hat{V}(x)}$$

Application to noisy Metropolis-Hastings (and comparison with Alquier et al.)

Rudolf et al. assume (i) K is geometrically ergodic, (ii) P satisfies a drift condition *i.e* $\delta \in (0,1)$ and $L > 0$

$$\forall\, x \in \mathsf{X}, \qquad KV(x) \leq \delta V(x) + L$$

then

$$W_d(\mu_0 K^n, \mu_0 \hat{K}^n) \leq C \frac{1-\rho^n}{1-\rho} \max\{\mu_0 V, L/(1-\delta)\} \sup_{x \in \mathsf{X}} \frac{\int Q(x, \mathrm{d}y) d(x,y) \, |\alpha(x,y) - \mathbb{E}_\eta\{\hat{\alpha}_\eta(x,y)\}|}{V(x)}$$

In contrast, Alquier et al. assume (i) K is uniformly ergodic and

$$\|\mu_0 K^n - \mu_0 \hat{K}^n\| \leq \lambda + C \frac{\rho^\lambda - \rho^n}{1-\rho} \sup_{x \in \mathsf{X}} \int Q(x, \mathrm{d}y) \mathbb{E}_\eta \, |\alpha(x,y) - \hat{\alpha}_\eta(x,y)|$$

Outline

Discussion

- Many statistical applications are incompatible with Metropolis-Hastings
- Alternative *exact* algorithms such as the exchange or the pseudo-marginal are in practice often disappointing
- Noisy MCMC methods are usually more efficient in practice (we have shown some such examples) but also more complicated to validate theoretically
- Replacing the acceptance ratio by a Monte Carlo estimate, one want to pick an estimator with a variance as small as possible (see Alquier et al.)
- Uniform ergodicity is a strong assumption for most MCMC applications (*e.g* unbounded & high dimensional) state space
- Ergodicity in the Wasserstein distance is appealing as the weighted nature of the distance (by the function V) allows to handle unbounded state space