Malaysian Journal of Mathematical Sciences 8(S): 103-108 (2014)
Special Issue: International Conference on Mathematical Sciences and Statistics 2013 (ICMSS2013)

**MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES**

Journal homepage: http://einspem.upm.edu.my/journal

# On the choice of Markov Kernels for Approximate Bayesian Computation

**Magid Maatallah**

*Department of Mathematics and Statistics,School of Business and Economics, Birkbeck, W85LF,University of London, UK*

*E-mail: msmatala@yahoo.com*

## ABSTRACT

Approximate Bayesian computation has emerged as a standard computational tool when dealing with the increasingly common scenario of completely intractable likelihood functions in Bayesian inference. We show that many common Markov chain Monte Carlo kernels used to facilitate inference in this setting can fail to be variance bounding, and hence geometrically ergodic, which can have consequences for the reliability of estimates in practice. We then prove that a recently introduced Markov kernel in this setting can be variance bounding and geometrically ergodic whenever its intractable Metropolis-Hastings counterpart is, under reasonably weak and manageable conditions. We indicate that the computational cost of the latter kernel is bounded whenever the prior is proper, and present indicative results on an example where spectral gaps and asymptotic variances can be computed Motivated by these considerations we study both the variance bounding and geometric ergodicity properties of a number of reversible kernels used for approximate Bayesian computations

Keywords: Approximate Bayesian computation; Markov chain Monte Carlo; Local adaptation.

## 1. INTRODUCTION

The Approximate Bayesian computation refers to branch of Monte Carlo methodology that utilizes the ability to simulate data according to a parameterized likelihood function in lieu of computation of that likelihood to perform approximate, parametric Bayesian inference. These methods have been used in an increasingly diverse range of applications since their inception in the context of population genetics (Tavare *et al.* (1997);

Pritchard *et al.* (1999)), particularly in cases where the likelihood function is either impossible or computationally prohibitive to evaluate.

We are in a standard Bayesian setting with data $y \in Y$ a parameter space $\Theta$, a prior $p:\Theta \to \mathbb{R}_+$ and for each $\theta \in \Theta$ likelihood $f_\theta : Y \to \mathbb{R}_+$ we assume Y is a metric space and consider the artificial likelihood.

$$f_\theta^\varepsilon(y) := V(\varepsilon)^{-1} \int_Y 1_{B_\varepsilon(x)}(y)\, f_\theta(x) dx = V(\varepsilon)^{-1} f_\theta(B_\varepsilon(y)) \tag{1}$$

This is commonly employed in approximate Bayesian computation. Here, $B_r(z)$ denotes a metric ball of radius r around z, $\mathrm{V}(\mathrm{r}) := \int_Y 1_{\mathrm{B_r(0)}}(\mathrm{x})\mathrm{dx}$ denotes the volume of a ball of radius r in Y and 1 is the indicator function. We adopt a slight abuse of notation by referring to densities as distributions, and where convenient, employ the measure-theoretic notation,

$$\mu(A) = \int_A \mu(d\lambda).$$

We consider situations in which both $\varepsilon$ and $y$ are fixed, and so define functions $h:\Theta \to [0,1]$ and $w:Y \to [0,1]$ by $h(\theta) := f_\theta(B_\varepsilon(y))$ and $w(x) := 1_{B_\varepsilon(x)}(y)$ to simplify the presentation. The value $h(\theta)$ can be interpreted as the the probability of "hitting" $B_\varepsilon(y)$ with a simple draw from $f_\theta$.

## 2. THE MARKOV KERNELS

In these sections we describe the algorithmic specification of the $\pi-$invariant Markov kernels under study. The algorithms specify how to sample from each kernel in each, a candidate $\vartheta$ is proposed according to a common proposal $q(\theta,.)$ and accepted or rejected, possibly along with other auxiliary variables, using simulations from likelihoods $f_\vartheta$ and $f_\theta$ we assume that for $\theta \in \Theta$, $q(\theta,.)$ and $p$ are densities with respect to a common dominating measure, e.g the Lebesgue or counting measures.

The first and most basic Markov kernel in this setting was proposed in Marjoram *et al*. (2003) and is a special case of a pseudo-marginal kernel

(Beaumont (2003)). Such kernels have been used in the context of approximate Bayesian computations in Del Moral *et al.* (2012) and involve on $\theta \times Y^N$ by additionally sampling auxiliary variables $z_{1:N} \sim f_\vartheta^{\otimes N}$ for a fixed $N \in \mathbb{N}$ .We denote kernels of this type for any $N$ by $P_{1,N}$ , and describe their simulation in algorithm 1.

**Algorithm 1.** To sample from $P_{1,N}\left((\theta, x_{1:N}),.\right)$

(1) Sample $\vartheta \sim q(\theta,.)$ and $z_{1:N} \sim f_\vartheta^{\otimes N}$ .

(2) With probability $1 \wedge \dfrac{p(\vartheta) q(\vartheta,\theta) \sum_{j=1}^{N} w(z_j)}{p(\vartheta) q(\vartheta,\theta) \sum_{j=1}^{N} w(x_j)}$, output $(\vartheta, z_{1:N})$.

Otherwise, output $(\theta, x_{1:N})$.

In Lee *et al.* (2012), two alternative kernels were proposed in this context, both of which evolve on $\Theta$. One denoted $P_{2,N}$ and described in Algorithm 2, is an alternative pseudo-marginal kernel that in addition to sampling $z_{1:N} \sim f_\vartheta^{\otimes N}$ , also samples auxiliary variables $x_{1:N-1} \sim f_\theta^{\otimes N-1}$. Detailed balance can be verified directly upon interpreting $S_z := \sum_{j=1}^{N} w(z_j)$ and $S_x := \sum_{j=1}^{N-1} w(x_j)$ as Binominal $(N, h(\vartheta))$ and binomial $(N-1, h(\theta))$ random variables respectively. The other kernel, denoted $P_3$ and described in Algorithm 3, also involves sampling according to $f_\theta$ and $f_\vartheta$ but does not sample a fixed number of auxiliary variables. This kernel also satisfies detailed balance (Lee (2012), Proposition 1).

**Algorithm 2.** To sample from $P_{2,N}(\theta,.)$

(1) Sample $\vartheta \sim q(\theta,.)$ and $x_{1:N-1} \sim f_\vartheta^{\otimes N-1}$ and $z_{1,N} \sim f_\vartheta^{\otimes N}$ .

(2) With probability $1 \wedge \dfrac{p(\vartheta) q(\vartheta,\theta) \sum_{j=1}^{N} w(z_j)}{p(\theta) q(\theta,\vartheta)\left[1 + \sum_{j=1}^{N-1} w(x_j)\right]}$, output $\vartheta$.

Otherwise output $\theta$ .

Because many of our positive results for $P_3$ are in relation to $P_{MH}$, the metropolis-Hastings kernel with proposal q, we provide the algorithmic specification of sampling from $P_{MH}$ in

**Algorithm 3.** To sample from $P_3(\theta,.)$

(1) Sample $\vartheta \sim q(\theta,.)$.

(2) With probability $1-\left(1\wedge\frac{p(\vartheta)q(\vartheta,\theta)}{p(\theta)q(\theta,\vartheta)}\right)$, stop and output $\theta$.

(3) For $i=1,2,....$ until $\sum_{j=1}^{i} w(z_j)+w(x_j)\geq 1$, sample $x_i \sim f_\theta$ and $z_i \sim f_\vartheta$. Set $N \leftarrow i$.

(4) If $w(Z_N)=1$, output $\vartheta$. Otherwise, output $\theta$.

Algorithm 4, we note that in the approximate Bayesian computation setting use of $P_{MH}$ is ruled out by assumption since $h$ cannot be computed and that the preceding kernels are, in some sense, "exact approximations" of $P_{MH}$.

**Algorithm 4**. To sample from $P_{MH}(\theta)$.

(1) Sample $\vartheta \sim q(\theta,.)$.

(2) With probability $\left(1\wedge\frac{p(\vartheta)h(\vartheta)q(\vartheta,\theta)}{p(\theta)h(\theta)q(\theta,\vartheta)}\right)$, output $\vartheta$. Otherwise, output $\theta$.

The kernels share a similar structure, and $P_{2,N}$, $P_3$ and $P_{MH}$ can each be written as

$$P(\theta,d\vartheta)=q(\theta,d\vartheta)\alpha(\theta,\vartheta)+\left(1-\int_{\Theta\setminus\{\theta\}} q(\theta,d\theta')\alpha(\theta,\theta')\right)\delta_\theta d\vartheta, \qquad (2)$$

Where only the function $\alpha(\theta,\vartheta)$ differs. $P_{1,N}$ can be represented similarly, modifications to account for its evolution on the extended space $\Theta\times Y^N$. The representation (4) is used extensively in our analysis and we have for $P_{2,N}$, $P_3$ and $P_{MH}$, respectively

$$\alpha_{2N}\left(\theta,\vartheta\int_{Y^N}\int_{Y^{N-1}}\left\{1-\wedge\frac{c(\vartheta,\theta)S_Z}{c(\theta,\vartheta)(1+S_X)}\right\}\right)f_\theta^{\otimes N-1}\left(d_{x_{1:N-1}}\right)f_\vartheta^{\otimes N}\left(d_{z_{1:N}}\right), \tag{3}$$

$$\alpha_3(\theta,\vartheta)=\left(1-\wedge\frac{C(\vartheta,\theta)}{C(\theta,\vartheta)}\right)\frac{h(\vartheta)}{h(\theta)+h(\vartheta)-h(\theta)h(\vartheta)}, \tag{4}$$

$$\alpha_{MH}(\theta,\vartheta)=1\wedge\frac{C(\vartheta,\theta)h(\vartheta)}{C(\theta,\vartheta)h(\theta)}, \tag{5}$$

where $c(\theta,\vartheta):=p(\theta)q(\theta,\vartheta)$ and (4) is obtained, e.g., in (Lee (2012)). Finally, we reiterate that the kernels satisfy detailed balance and are therefore reversible.

## 3. THEORITICAL PROPERTIES

We assume that $\Theta$ is a metric space, and that $H:=\int p(\theta)h(\theta)d(\theta)$ satisfies $H\in(0,\infty)$ so $\pi$ is well defined. We allow $p$ to be improper, i.e. for $\int p(\theta)d\theta=1$ to be unbounded, but otherwise assume $p$ is normalized, i.e. $\int p(\theta)d\theta=1$. We define the collection of local proposals to be

$$Q:=\left\{q:\forall\delta\geq 0,\exists r\in(0,\infty),\forall\theta\in\Theta,q\left(\theta,B_r^c(\theta)\leq\delta\right)\right\}, \tag{6}$$

which encompasses a broad number of common choices in practice. We denote by $\upsilon$ and $\varsigma$ the collections of variance bounding and geometrically ergodic kernels, respectively, noting that $\varsigma\subset\upsilon$. In our analysis, we make use of of the following conditions.

$(C_1)$ (i) $q\in Q$.

(ii) $\forall r\geq 0,\ \pi\left(B_r^c(0)\right)\geq 0$.

(iii) $\forall\delta\geq 0,\ \exists\nu\geq 0,\ \sup_{\theta\in B_\nu^c(0)}h(0)\leq\delta$.

$(C_2)$ (i) $q\in Q$.

(ii) $\forall K\geq 0,\exists M_K\in(0,\infty),\forall(\theta,\vartheta)\in\Theta\times B_K(\theta):\pi(\theta)q(\theta,\vartheta)\wedge$

$$\pi(\vartheta)q(\vartheta,\theta)\geq 0,\ \text{either}\ \ \frac{h(\vartheta)}{h(\theta)}\in\left[M_K^{-1},M_K\right]\ \ \text{or}$$

$$\frac{c(\vartheta,\theta)}{c(\theta,\vartheta)}\in\left[M_K^{-1},M_K\right].$$

$$(C_3)\quad \exists M\in(0,\infty),\forall(\theta,\vartheta)\in\Theta^2:\pi(\theta)q(\theta,\vartheta)\wedge\pi(\vartheta)q(\vartheta,\theta)\geq 0,$$
$$\text{either}\ \frac{h(\vartheta)}{h(\theta)}\in\left[M_K^{-1},M_K\right]\ \text{or}\ \frac{c(\vartheta,\theta)}{c(\theta,\vartheta)}\in\left[M^{-1},M\right].$$

## 4. DISCUSSION

In this our analysis suggests that $P_3$ may be geometrically ergodic and or variance bounding in a wide variety of situations where kernels $P_{1,N}$ and $P_{2,N}$ are not. In practice, $(C_2)$ can be verified and used to inform prior and proposal choice to ensure that $P_3$ systematically inherits these properties from $P_{MH}$. Of course variance bounding or geometric ergodicity of $P_{MH}$ is typically impossible to verify in the approximate Bayesian computation setting due to the unknown nature of $f_\theta$. However, a prior with regular contours will ensure that $P_{MH}$ is geometrically ergodic if $f_\theta$ decays super-exponentially and also has a regular contours. In addition, $(C_2)$ and $(C_3)$ are stronger than necessary but tighter conditions are likely to be complicated and may require case-by-case treatment.

## REFERENCES

Lee, A., Andrieu, C. and Doucet, A. (2012). Discussion of paper by P. Fearnhead and D. Prangle. *J. R. Statist. Soc*. **B**(74): 419-474.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation. *J. R. Statist. Soc.* **B**(74): 419-474.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov chains and stochastic stability.* Cambridge University Press.

Roberts, G. O. and Rosenthal, J. S. (2006).Variance bounding Markov chains. *Ann. Appl. Prob.* **18**: 1201.