The Containment Condition and AdapFail algorithms

Krzysztof Łatuszyński and Jeffrey S. Rosenthal

K. Łatuszyński
Department of Statistics
University of Warwick
CV4 7AL, Coventry, UK
e-mail: latuch@gmail.com

J. S. Rosenthal
Department of Statistics
University of Toronto
Toronto, Ontario, Canada, M5S 3G3
e-mail: jeff@math.toronto.edu

(Version of July 6, 2013)

Abstract: This short note investigates convergence of adaptive MCMC algorithms, i.e. algorithms which modify the Markov chain update probabilities on the fly. We focus on the Containment condition introduced in [RR07]. We show that if the Containment condition is *not* satisfied, then the algorithm will perform very poorly. Specifically, with positive probability, the adaptive algorithm will be asymptotically less efficient then *any* nonadaptive ergodic MCMC algorithm. We call such algorithms `AdapFail`, and conclude that they should not be used.

AMS 2000 subject classifications: Primary 60J05, 65C05.
Keywords and phrases: Markov chain Monte Carlo, adaptive MCMC, Containment condition, ergodicity, convergence rates.

1. Introduction

Markov chain Monte Carlo (MCMC) algorithms are used to sample from complicated probability distributions. They proceed by simulating an ergodic Markov chain with transition kernel P and stationary distribution of interest, say π. Unlike in the case of iid Monte Carlo, the MCMC output

$$X_0, X_1, ..., X_n, ... \tag{1}$$

is a correlated sample. Nevertheless, due if the Markov chain is ergodic (i.e., converges in distribution to π), then the asymptotic validity is retained under appropriate conditions (see e.g. [MT09, RR04]). In particular, for M large enough, the subsampled random variables

$$X_M, X_{2M}, ..., X_{nM}, ... \tag{2}$$

are approximately independent draws from the target distribution π. For the MCMC-based statistical inference to be reliable, it is essential to design algorithms that mix quickly , i.e. for which the asymptotic iid property in (2) holds with reasonably small M. (For estimation purposes, the whole sample (1) should still be used; see e.g. [Gey92].)

In a typical MCMC setting, the algorithm is determined by a Markov chain transition kernel P_θ, where $\theta \in \Theta$ is a high dimensional tuning parameter, e.g. the covariance matrix of a Random Walk Metropolis proposal [RGG97, RR01], or the vector of Random Scan Gibbs Sampler selection probabilities [ŁRR13]. Usually the parameter space Θ is large, and for "good" values of θ, the iterates P_θ^n will converge quickly to π as n increases, resulting in small M in (2). However, such "good" values are often very difficult to find, and for most values of θ the iterates P_θ^n will converge arbitrary slowly.

Since a good θ is difficult to find manually, the idea of adaptive MCMC was introduced [GRS98, HST01] to enable the algorithm to learn "on the fly", and redesign the transition kernel during the simulation as more and more information about π becomes available. Thus an adaptive MCMC algorithm would apply the transition kernel P_{θ_n} for obtaining X_n from X_{n-1}, where the choice of the tuning parameter θ_n at the n^{th} iteration is itself a random variable which may depend on the whole history $X_0, X_1, ..., X_{n-1}$ and on θ_{n-1}. When using adaptive MCMC, one hopes that the adaptive parameter θ_n will settle on "good" values, and that the adaptive algorithm will inherit the corresponding good convergence properties.

Unfortunately, since adaptive algorithms violate the Markovian property, they are inherently difficult to analyse theoretically. Whereas the interest in adaptive MCMC is fuelled by some very successful implementations for challenging problems [RR09, AT08, RBR10, GŁS13, GK08, SOL+12], many seemingly reasonable adaptive MCMC algorithms are provably transient or converge to a wrong probability distribution [AR05, BRR11, ŁRR13, Łat12]. Thus, the theoretical foundations of adaptive MCMC are a very important topic which is still under active development.

One general and relatively simple approach to analysing adaptive MCMC algorithms was presented in [RR07], which showed that the two properties of *Diminishing Adaptation* and *Containment* were sufficient to guarantee that an adaptive MCMC algorithm would converge asymptotically to the correct target distribution (at some rate). While the Diminishing Adaptation property is fairly standard and can be easily controlled by the user, the Containment property is more subtle and can be challenging to verify (see e.g. [BRR10]). This leads to the question of how important or useful the Containment condition actually is, especially since it is known (see e.g. [FMP11]) that Containment is not a necessary condition for the ergodicity of an adaptive MCMC algorithm.

The purpose of this short note is to show that if Containment does not hold, then the adaptive algorithm will perform very poorly. Specifically, with positive probability the adaptive algorithm will be asymptotically less efficient then *any* nonadaptive MCMC algorithm. In effect, the approximate iid property in (2) will be violated for *any* finite M. We call such algorithms `AdapFail`, and

conclude that they should not be used. In particular, this argues that the Containment condition is actually a reasonable condition to impose on adaptive MCMC algorithms, since without it they will perform so poorly as to be unusable.

This paper is structured as follows. In Section 2, we define and characterise the class of `AdapFail` algorithms. In Section 3, we relate the `AdapFail` property to the Containment condition. In Section 4, we present a very simple example to illustrate our results.

2. The class of `AdapFail` algorithms

We first introduce necessary notation; see e.g. [MT09, RR04, RR07] for more complete development related to Markov chains and adaptive MCMC. Let P_θ, parametrized by $\theta \in \Theta$, be a transition kernel of a Harris ergodic Markov chain on $(\mathcal{X}, \mathcal{F})$ with stationary distribution π. Thus for all $x \in \mathcal{X}$ and $\theta \in \Theta$ we have $\lim_{n\to\infty} \|P_\theta^n(x,\cdot) - \pi(\cdot)\| = 0$, where $\|\nu(\cdot) - \mu(\cdot)\| := \sup_{A\in\mathcal{F}} |\nu(A) - \mu(A)|$ is the usual total variation norm. We shall also use the "ε convergence time function" $M_\varepsilon : \mathcal{X} \times \Theta \to \mathbb{N}$ defined as

$$M_\varepsilon(x,\theta) := \inf\{n \geq 1 : \|P_\theta^n(x,\cdot) - \pi(\cdot)\| \leq \varepsilon\}. \tag{3}$$

Let $\{(X_n, \theta_n)\}_{n=0}^\infty$ be a corresponding adaptive MCMC algorithm, where X_n is updated from X_{n-1} using P_{θ_n} for *some* Θ-valued random variable θ_n (which might depend on the chain's history and on θ_{n-1}). For the adaptive algorithm, denote the marginal distribution at time n by

$$A^{(n)}((x,\theta), B) := \mathbb{P}(X_n \in B | X_0 = x, \theta_0 = \theta), \tag{4}$$

and say that the algorithm is *ergodic* for starting values x and θ if

$$\lim_{n\to\infty} \|A^{(n)}((x,\theta),\cdot) - \pi(\cdot)\| \quad = \quad 0. \tag{5}$$

Similarly let the "ε convergence time function" for the adaptive case be

$$M_\varepsilon^A(x,\theta) := \inf\{n \geq 1 : \|A^{(n)}((x,\theta),\cdot) - \pi(\cdot)\| \leq \varepsilon\}. \tag{6}$$

In both cases the function $M_\varepsilon(x,\theta)$ has the same interpretation: it is the number of iterations that the algorithm must take to be within ε of stationarity.

We are now ready to define the class of `AdapFail` algorithms.

Definition 2.1. Let $\{(X_n, \theta_n)\}_{n=0}^\infty$ evolve according to the dynamics of an adaptive MCMC algorithm $\mathcal{A}$, with starting values $X_0 = x^*$ and $\theta_0 = \theta^*$. We say that $\mathcal{A} \in$ `AdapFail` if there is $\varepsilon_{AF} > 0$ such that

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}\big(M_{\varepsilon_{AF}}^A(X_n,\theta_n) > M \mid X_0 = x^*, \theta_0 = \theta^*\big) \;=:\; \delta_{AF} \;>\; 0\,. \tag{7}$$

Remark 2.2. Intuitively, (7) says that the convergence times of the adaptive algorithm will be larger than any fixed value M, i.e. that the algorithm will converge arbitrarily slowly and thus perform so poorly as to be unusable.

Remark 2.3. In our experience, the inner limit in (7) will typically exist, so that $\limsup_{n\to\infty}$ can be replaced by $\lim_{n\to\infty}$ there (and similarly in the related expressions below).

Remark 2.4. For the probabilities in (7) to make sense, the function M_ε^A needs to be measurable. This follows from the Appendix of [RR97]. Moreover, if the inner limit in (7) is denoted as $\delta_{AF}(M)$, then this sequence is positive and non-increasing as a function of M, and will thus converge to δ_{AF} as $M \to \infty$.

Remark 2.5. To obtain the approximate iid property of the $\{X_n\}$ in (2), we want the distribution of $X_{(n+1)M}$ conditionally on the value of X_{nM} to be within ε of the stationary measure, i.e.

$$\big\|\mathcal{L}\big(X_{(n+1)M} \mid X_{nM}\big) - \pi\big\| \leq \varepsilon. \tag{8}$$

Being an `AdapFail` algorithm means that for any fixed $0 < \varepsilon \leq \varepsilon_{AF}$ and some fixed $\delta_{AF} > 0$, we are infinitely often in a regime where (8) is violated for *any* finite M, with probability at least δ_{AF}, further illustrating its poor performance.

The following two results shed additional light on the `AdapFail` class.

Proposition 2.6. *Any ergodic nonadaptive MCMC algorithm P_θ is not in* ***`AdapFail`****.*

Proof. For a nonadaptive chain, the quantity M_ε^A in (7) becomes M_ε and $\theta^* = \theta$. For arbitrary $\varepsilon > 0$ and $\delta > 0$, we shall show that $\delta_{AF} < 2\delta$, from which it follows that $\delta_{AF} = 0$. Indeed, first find n_0 such that $\|P_\theta^{n_0}(x^*, \cdot) - \pi(\cdot)\| < \delta$, and then find M_0 such that $\pi(\{x : M_\varepsilon(x,\theta) \leq M_0\}) > 1 - \delta$. Then for every $n \geq n_0$ and every $M \geq M_0$, we can write

$$\mathbb{P}\big(M_\varepsilon(X_n,\theta) > M \mid X_0 = x^*\big) \quad \leq \quad \delta + \pi(\{x : M_\varepsilon(x,\theta) > M\}) \quad < \quad 2\delta.$$

The result follows. □

Theorem 2.7. *For an algorithm $\mathcal{A}$ the following conditions are equivalent.*

(i) $\mathcal{A} \in$ ***`AdapFail`****.*

(ii) there are $\varepsilon > 0$ and $\delta > 0$ such that for all $x \in \mathcal{X}$, $\theta \in \Theta$, and $K > 0$,

$$\limsup_{n\to\infty} \ \mathbb{P}\big(M_\varepsilon^A(X_n,\theta_n) > K M_\varepsilon(x,\theta) \mid X_0 = x^*, \theta_0 = \theta^*\big) \ \geq \ \delta.$$

(iii) there are $\varepsilon > 0$ and $\delta > 0$ such that for all $\theta \in \Theta$, $K > 0$, and $y^* \in \mathcal{X}$,

$$\limsup_{n\to\infty} \ \mathbb{P}\big(M_\varepsilon^A(X_n,\theta_n) > K M_\varepsilon(Y_n,\theta) \mid X_0 = x^*, \theta_0 = \theta^*, Y_0 = y^*\big) \ \geq \ \delta,$$

where $\{Y_n\}$ is a Markov chain which follows the dynamics P_θ and is independent of the adaptive process $\{X_n\}$.

Moreover, in (ii) and (iii) we can take $\delta = \delta_{AF}$.

Proof. The implication $(i) \Rightarrow (ii)$ with $\delta = \delta_{AF}$ and $\varepsilon = \varepsilon_{AF}$ is immediate. To verify $(ii) \Rightarrow (iii)$, fix $\delta^* > 0$ such that $\|P^n(y^*, \cdot) - \pi(\cdot)\| \leq \delta_0$ for infinitely many n. Then for fixed θ, K, and y^*, compute

$$\begin{aligned}
&\mathbb{P}\big(M^A_{\varepsilon_{AF}}(X_n, \theta_n) > K M_{\varepsilon_{AF}}(Y_n, \theta) \mid X_0 = x^*, \theta_0 = \theta^*, Y_0 = y^*\big) \\
&\quad = \int_{\mathcal{X}} \mathbb{P}\big(M^A_{\varepsilon_{AF}}(X_n, \theta_n) > K M_{\varepsilon_{AF}}(x, \theta) \mid X_0 = x^*, \theta_0 = \theta^*\big) P^n(y^*, dx) \\
&\quad \geq \int_{\mathcal{X}} \mathbb{P}\big(M^A_{\varepsilon_{AF}}(X_n, \theta_n) > K M_{\varepsilon_{AF}}(x, \theta) \mid X_0 = x^*, \theta_0 = \theta^*\big) \pi(dx) \;-\; \delta^* \\
&\quad \to \;\; \Delta_{AF} - \delta^* \;\; \geq \;\; \delta_{AF} - \delta^*,
\end{aligned}$$

where the last limit is along a sequence $n \to \infty$ such that $\|P^n(y^*, \cdot) - \pi(\cdot)\| \leq \delta_0$; the limit must exist by dominated convergence theorem. Hence (iii) follows with $\delta = \delta_{AF}$, since δ^* was arbitrary. For $(iii) \Rightarrow (i)$, notice that $M_{\varepsilon_{AF}}(Y_n, \theta) \geq 1$, so (iii) gives

$$\limsup_{n \to \infty} \; \mathbb{P}\big(M^A_{\varepsilon_{AF}}(X_n, \theta_n) > K \mid X_0 = x^*, \theta_0 = \theta^*\big) \;>\; \delta_{AF}, \quad \text{for every } K > 0.$$

The result follows by taking $K \to \infty$. □

Remark 2.8. Condition (iii) has the interpretation that if we run the adaptive algorithm $\{X_n\}$ and a nonadaptive $\{Y_n\}$ independently on two computers next to each other, and monitor the ε convergence time of both algorithms, then as the simulation progress, with probability at least δ, the ε convergence time of the adaptive algorithm will infinitely often be bigger by an arbitrarily large factor K, i.e. $\{X_n\}$ will be arbitrarily worse than $\{Y_n\}$ (no matter how bad are the tuning parameters θ and starting point Y_0 for $\{Y_n\}$).

3. Relation to the Containment condition

The following condition was introduced in [RR07] as a tool to analyse adaptive MCMC algorithms:

Definition 3.1 (Containment Condition)**.** The algorithm $\mathcal{A}$ with starting values $X_0 = x^*$ and $\theta_0 = \theta^*$ satisfies Containment, if for all $\varepsilon > 0$ the sequence $\{M_\varepsilon(X_n, \theta_n)\}_{n=0}^{\infty}$ is bounded in probability.

It is augmented by the usual requirement of Diminishing Adaptation:

Definition 3.2 (Diminishing Adaptation)**.** The algorithm $\mathcal{A}$ with starting values $X_0 = x^*$ and $\theta_0 = \theta^*$ satisfies Diminishing Adaptation, if

$$\lim_{n \to \infty} D_n = 0 \quad \text{in probability, where} \quad D_n := \sup_{x \in \mathcal{X}} \|P_{\theta_{n+1}}(x, \cdot) - P_{\theta_n}(x, \cdot)\|.$$

Containment has been extensively studied in [RR07] and [BRR11] and verified for large classes of adaptive MCMC samplers (c.f. also [RR09, ŁRR13]). Together with Diminishing Adaptation, it guarantees ergodicity. As illustrated in the next section, it is *not* a necessary condition. However, it still turns out to be an appropriate condition to require, due to the following result.

Theorem 3.3. *Assume the Diminishing Adaptation is satisfied. Then the Containment condition does not hold for $\mathcal{A}$ if and only if $\mathcal{A} \in$ `AdapFail`.*

Proof. The proof utilises a construction similar to the coupling proof of Theorem 1 of [RR07] (see also [RR13]). First, by the Diminishing Adaptation property, for any fixed $\delta_c > 0$, $\varepsilon_c > 0$, and integer $M \geq 1$, we can choose n big enough that

$$\mathbb{P}\Big(\bigcup_{k=1}^{M} \{D_{n+k} > \frac{\varepsilon_c}{2M^2}\}\Big) \;\leq\; \frac{\delta_c}{2}. \tag{9}$$

Now, on the set $\bigcap_{k=1}^{M}\{D_{n+k} \leq \frac{\varepsilon_c}{2M^2}\}$ for transition kernels $P_{\theta_n}, P_{\theta_{n+1}}, ..., P_{\theta_{n+M}}$, by the triangle inequality we have

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \Big\| \big(\prod_{k=0}^{M} P_{\theta_{n+k}}\big)(x,\cdot) - P_{\theta_n}^{M}(x,\cdot)\Big\| \;\leq \qquad (10)\\
&\quad\leq \sum_{k=1}^{M} \sup_{x\in\mathcal{X}} \Big\| \Big(\big(\prod_{i=0}^{k+1} P_{\theta_{n+i}}\big) P_{\theta_n}^{M-k-1}\Big)(x,\cdot) - \Big(\big(\prod_{i=0}^{k} P_{\theta_{n+i}}\big) P_{\theta_n}^{M-k}\Big)(x,\cdot)\Big\| \\
&\quad\leq \sum_{k=1}^{M} (k+1)\frac{\varepsilon_c}{2M^2} \;=\; \frac{M+1}{4M}\varepsilon_c \;<\; \frac{\varepsilon_c}{2}.
\end{aligned}$$

Consequently we conclude that for n large enough,

$$\mathbb{P}\Big(\text{LHS of (10)} \;<\frac{\varepsilon_c}{2}\Big) \;>\; 1 - \frac{\delta_c}{2}. \tag{11}$$

For the "only if" part of the theorem, note that if Containment does not hold, then for the adaptive algorithm in question, there is $\varepsilon_c > 0$ and $\delta_c > 0$ such that

$$\forall M,\, n_0,\; \exists n > n_0 \quad \text{s.t.} \quad \mathbb{P}(M_{\varepsilon_c}(X_n, \theta_n) > M) > \delta_c. \tag{12}$$

By (11), we obtain

$$\forall M,\, n_0,\; \exists n > n_0 \quad \text{s.t.} \quad \mathbb{P}(M^{A}_{\varepsilon_c/2}(X_n, \theta_n) > M) > \frac{\delta_c}{2}. \tag{13}$$

which implies the `AdapFail` condition with $\varepsilon_{AF} \geq \varepsilon_c/2$ and $\delta_{AF} \geq \delta_c/2$.

The proof for the "if" part of the theorem is essentially the same. From (7) and (11), one obtains

$$\forall M,\; \exists n_0 \quad \text{s.t.} \quad \forall n \geq n_0, \qquad \mathbb{P}(M_{\varepsilon_c}(X_n, \theta_n) > M) > \delta_c, \tag{14}$$

with $\varepsilon_c \geq \varepsilon_{AF}/2$ and $\delta_c \geq \delta_{AF}/2$. Condition (14) then implies (12). □

4. A very simple example

In this section, we analyse a very simple example of an adaptive algorithm, to illustrate our results about `AdapFail`.

Example 4.1. Consider the toy example from [FMP11] with state space $\mathcal{X} = \{0, 1\}$ and stationary distribution $\pi = (1/2, 1/2)$, with Markov transition kernels

$$P_\theta = \begin{pmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{pmatrix}.$$

Suppose the n^{th} iteration of the Markov chain uses kernel P_{θ_n} (independent of the chain's past history), where $\theta_n > 0$ and $\sum_n \theta_n = \infty$ but $\theta_n \to 0$ (e.g. $\theta_n = 1/n$). Since the θ_n converges, clearly Diminishing Adaptation is satisfied. On the other hand, as $\theta \to 0$, $M_\epsilon(x, \theta) \to \infty$. Hence, this adaptive algorithm does *not* satisfy Containment. So, by the above theorems, this algorithm converges more slowly than any fixed non-adaptive algorithm. But since $\sum_n \theta_n = \infty$, this algorithm is still ergodic [FMP11]. We thus have a (very simple) example of an adaptive algorithm which is ergodic, but is nevertheless in AdapFail and has very poor convergence properties. (A similar result presents if instead $\theta_n \to 1$ with $\sum_n (1 - \theta_n) = \infty$.)

5. Acknowledgements

KŁ acknowledges funding from CRISM and other grants from EPSRC. JSR acknowledges funding from NSERC of Canada. We thank Gersende Fort and Gareth O. Roberts for helpful discussions.

References

[AR05] Y.F. Atchadé and J.S. Rosenthal. On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828, 2005.

[AT08] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

[BRR10] Y. Bai, G.O. Roberts, and J.S. Rosenthal. On the containment condition for adaptive Markov chain Monte Carlo algorithms. *Preprint*, 2010.

[BRR11] Y. Bai, G.O. Roberts, and J.S. Rosenthal. On the containment condition for adaptive markov chain monte carlo algorithms. *Advances and Applications in Statistics*, 21(1):1–54, 2011.

[FMP11] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting markov chain monte carlo algorithms. *The Annals of Statistics*, 39(6):3262–3289, 2011.

[Gey92] Charles J Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483, 1992.

[GK08] P. Giordani and R. Kohn. Efficient bayesian inference for multiple change-point and mixture innovation models. *Journal of Business and Economic Statistics*, 26(1):66–77, 2008.

[GŁS13] J.E. Griffin, K. Łatuszyński, and M.F.J. Steel. Individual adaptation: an adaptive MCMC scheme for variable selection problems. *submitted*, 2013.

[GRS98] W.R. Gilks, G.O. Roberts, and S.K. Sahu. Adaptive markov chain monte carlo through regeneration. *Journal of the American Statistical Association*, 93(443):1045–1054, 1998.

[HST01] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[Łat12] K. Łatuszyński. A path stability condition for adaptive mcmc. *in preparation*, 2012.

[ŁRR13] K. Łatuszyński, G.O. Roberts, and J.S. Rosenthal. Adaptive Gibbs samplers and related MCMC methods. *Ann. Appl. Probab.*, 23(1):66–98, 2013.

[MT09] S.P. Meyn and R.L Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, 2009.

[RBR10] S. Richardson, L. Bottolo, and J.S. Rosenthal. Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics*, 9, 2010.

[RGG97] G.O. Roberts, A. Gelman, and W.R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

[RR97] G.O. Roberts and J.S. Rosenthal. Geometric ergodicity and hybrid Markov chains. *Electron. Comm. Probab*, 2(2):13–25, 1997.

[RR01] G.O. Roberts and J.S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, 2001.

[RR04] G.O. Roberts and J.S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[RR07] G.O. Roberts and J.S. Rosenthal. Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability*, 44(2):458, 2007.

[RR09] G.O. Roberts and J.S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[RR13] Gareth O Roberts and Jeffrey S Rosenthal. A note on formal constructions of sequential conditional couplings. *Statistics & Probability Letters, to appear*, 2013.

[SOL+12] Antti Solonen, Pirkka Ollinaho, Marko Laine, Heikki Haario, Johanna Tamminen, and Heikki Järvinen. Efficient mcmc for climate model parameter estimation: Parallel adaptive chains and early rejection. *Bayesian Analysis*, 7(3):715–736, 2012.