On the Ergodicity of Adaptive MCMC Algortihms

Eero Saksman (Helsinki)

MCQMC2010

Warsaw, 19.8.2010

The aim of the talk

- Recall basic adaptive MCMC.
- Overview the development of the general understanding of ergodicity of adaptive MCMC.
- If time allows, to decribe in more detail a result (joint with M. Vihola) on the ergodicity of the adaptive Metropolis algorithm.

Warning !

- Our overview is very partial!
- Most technicalities will be downplayed or not explained for simplicity
  $\Longrightarrow$ even statements of some results are only approximate!
- We will concentrate on algorithms with 'diminishing adaptation', and mainly consider implications of known convergence results for the Adaptive Metropolis algorithm.

MCMC algorithms: Ergodicity

Assume that $\pi(x)$ is the density of a given probability distribution on $\mathbb{R}^n$ and we need to simulate π (or some integrals $\int_{\mathbb{R}^d} f(x)\pi(x)\,dx$).

In **MCMC** one solves this problem by constructing an (algorithmic) Markov chain X_n on $\mathbb{R}^n$ such that (X_n) properly *ergodic* with respect to π. The simulation is then simple: start the chain from an arbitrary initial point and let it run.

Correct simulation requires that chain has the right ergodic properties.
A minimal requirement:

$$X_k \sim \pi \qquad X_{k+1} \sim \pi.$$

Equivalently, if $P = P(x,dy)$ stands for the transition probability kernel of the chain (X_n)

$$\pi P = \pi.$$

Egodicity (continued)

One needs loss of memory:

$$\lim_{n\to\infty} \|P^n(x,\cdot) - \pi\| = 0 \quad \text{for all } x.$$

This is not enough, correct simulation requires

(E) $$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \int_{\mathbb{R}^n} f(x)\, dx.$$

for all ('nice', e.g. bounded) $f : \mathbb{R}^n \to \mathbb{R}$ (preferably almost surely).

If (E) holds, we say that the algorithm is **ergodic**.

For standard MCMC this is most often automatically valid

Ramark: The adaptive chains we will consider are not Markov chains !
One has to be more careful in using familiar concepts, e.g. when deducing (E) from other properties of the chain.

Need for adaptive MCMC

Recall **the Metropolis algorithm**: let $X_0 \equiv x_0$ and iterate for $n \geq 1$,

1. simulate $Y_n = X_{n-1} + U_n$, where U_n is an independent random variable distributed according to some symmetric **proposal distribution** q, and
2. with probability $\min\{1, \pi(Y_n)/\pi(X_{n-1})\}$, the proposal is accepted and $X_n = Y_n$; otherwise the proposal is rejected and $X_n = X_{n-1}$.

We chall consider only the case where the proposal q is a Gaussian. Let q have covariance C and denote by P_C the corresponding transition kernel of the Metropolis Chain.

• The efficiency of the algorithm **depends heavily** on the choice of the proposal, i.e. on the choice of the covariance C !

• In many applications one performs large number of simulations with varying targets

$\Longrightarrow$ Want **automaticed tuning** of C, i.e. **adaptive algorithms**

The Adaptive Metropolis (AM) algorithm

[Haario, S., Tamminen 2001] defined the **AM algorithm** so that the proposal is not fixed, but it learns from the full history of the chain, trying to 'adapt' to π:

1. After X_n is determined, compute C_n from the empirical covariance:

$$C_n = \varepsilon Id + \gamma_d Cov(X_1, X_2, \ldots, X_n)$$

2. Then compute compute X_{n+1} by a standad Metropolis step by applying Gaussian proposal with covariance C_n.
3. Go to step 1.

- There is a simple recursion formula for the covariance that allows computationally efficient implementation. Often one applies a burn in period.
- AM appears to work in a quite robust manner, also in fairly high dimensions. Limitations: e.g., very high dimensions (?) or significant multimodality.
- AM and variants of it are used increasingly often in various real applications.
- Prior to AM there existed other adaptive MCMC algorithms, e.g. the one by [Gilks, Roberts, Sahu,1998] adapted when returning to a fixed atom.

The first ergodicity results for AM

Theorem. ([H,S, T,2001]) *Assume that* π *is compactly supported and bounded from above. Then the AM-chain is ergodic.*

Elements of the proof:

- The conditions on π (together with the condition $\varepsilon > 0$) make sure that C_n stays bounded from below and above. This implies that the n:th step kernels P_{C_n} stay **uniformly ergodic**:

$$\|P_{C_n}^k(x,\cdot) - \pi\| \leq C\lambda_0^k \quad \text{for all } k \quad (\text{here } \lambda_0 \in (0,1))$$

- By the recursion $\|C_{n+1} - C_n\| \leq c/n$ ('**diminishing adaptation**')
- **Local coupling**: For $1 << k << n_0$ the part of the chain

$$X_{n_0}, \ldots, X_{n_0+k}$$

is compared to to the Markovian approximation

$$\widetilde{X}_{n_0}, \ldots, \widetilde{X}_{n_0+k},$$

where one applies $C_n = C_{n_0}$ for $n_0 \leq n < n_0 + k$.

The first ergodicity results for AM (continued)

However, the above result was quite restricted. Especially, it left open two important questions:

- What happens for **targets with non-compact support** ?
- How fast is the convergence, is CLT valid?

REMARK:

- It is quite **important to understand ergodicity for adaptive strategies.** Simple examples show that some natural looking adaptation schemes either converge to a wrong limit distribution, or 'do not converge' at all !

General adaptation schemes, connection to stochastic approximation

[Andrieu, Robert, 2001, technical report]: more **general adaptation schemes**, where one adapts any parameters of the MCMC chain while keeping '**diminishing adaptation**' property.

Consider an adaptive MCMC chain (X_n), and the chain of adaptation parameters (C_n) evolving in $\mathbb{R}^{d'}$. One way to introduce more general formulation is to consider the recursion

1. $X_{n+1} \sim P_{C_n}(X_n, \cdot)$
2. $C_{n+1} = C_n + \eta_{n+1} H(C_n, X_{n+1})$.

Here H is an **adaptation function** and $\eta_1 \geq \eta_2 \ldots \geq \eta_n \to_{n\to\infty} 0$ is a decreasing sequence of **adaptation step sizes**.

$\Longrightarrow$ connection to **stochastic approximation**. This can be helpful in proving convergence results, or in devicing adaptation schemes.

Formulating AM inside the general scheme

Denote (**from now on**) by $\widetilde{C}_n$ the covariance of the Gaussian proposal at step n. Let $M_n = (1/(n+1))\sum_{k=0}^{n} X_k$ be the empirical mean. Write

$$C_n = (M_n, \widetilde{C}_n).$$

Then the original AM (essentially) corresponds to the choices

$$H((\widetilde{c}, m), x) := \begin{bmatrix} x - m \\ (x-m)(x-m)^T - \widetilde{c} + \kappa I \end{bmatrix}$$

and

$$\eta_n = (n+1)^{-1} \quad \text{for } n \geq 1.$$

A generalization of the original proof

- [Atchade, Rosenthal, 2005] uses the original proof strategy, including the use of mixingales.
- Despite the nice results, one cannot easily get verifiable results for e.g. AM because of the **implicite nature of the assumptions** made.

Non-compactly supported targets, modified AM

Recall that the kernel P is **geometrically ergodic** if there is a function K, and $\lambda_0 \in (0,1)$ so that

$$\|P^n(x,\cdot) - \pi\| \leq K(x)\lambda_0^n \quad \text{for all } x.$$

A standard Metropolis (say, with a Gaussian proposal), is known to be ([Jarner and Hansen(2000)]) geometrically ergodic if (JH_ρ) holds for $\rho = 1$, where (besides minor extra conditions)

(JH_ρ) $\quad \lim_{r\to\infty} \sup_{\|x\|\geq r} \frac{x}{\|x\|^\rho} \cdot \nabla \log \pi(x) = -\infty$ and

$$\lim_{r\to\infty} \sup_{\|x\|\geq r} \frac{x}{\|x\|} \cdot \frac{\nabla\pi(x)}{\|\nabla\pi(x)\|} < 0.$$

Theorem. ([Andrieu, Moulines, 2006]) *Assume that π satisfies (JH_1).Then the AM-chain,* **modified with back-projections***, is ergodic and satisfies CLT.*

Here **back-projections** return the covariance to a fixed value if it goes over certain increasing threshold value, that is increased after each such instance.

Corollary: CLT for the unmodified AM in case of compactly supported targets.

Elements of the proof

[Andrieu, Moulines, 2006] uses careful quantitative estimates for geometrically ergodic chains together with stochastic approximation theory. The applicability of Poisson kernel techniques is observed, especially one makes use of the important decomposition (due, in another context, to Benvetiste, Metivier and Prioret ?):

$$\sum_{j=1}^{k} \left[f(X_j) - \pi(f) \right] = M_k + R_k^{(1)} + R_k^{(2)},$$

where M_k is the **martingale** partial sum

$$M_k := \sum_{j=1}^{k} \left[\widehat{f}_{C_{j-1}}(X_j) - P_{C_{j-1}} \widehat{f}_{C_{j-1}}(X_{j-1}). \right]$$

Here $\widehat{f}_C$ satisfies the Poisson equation

$$\widehat{f}_C - P_C \widehat{f}_C = f - \pi(f).$$

The contribution from (M_k) can be taken care by the martingale limit theorem.

The residual terms

The remaining terms are

$$\begin{aligned} R_k^{(1)} &:= \sum_{j=1}^{k} \left[\widehat{f}_{C_j}(X_j) - \hat{f}_{C_{j-1}}(X_j) \right] \\ R_k^{(2)} &:= P_{C_0} \hat{f}_{C_0}(X_0) - P_{C_k} \hat{f}_{C_k}(X_k). \end{aligned}$$

Here $R_k^{(2)}$ is (relatively) harmless.

$R_k^{(2)}$ can be thought of measuring the **contribution from adaptation** , and usually it can be shown to be very small compared to the martingale term !

Drawback: back projections present an unwelcome modification.

Coupling methods: diminishing adaptation and containment

The paper [Roberts, Rosenthal, 2007] applies **coupling methods** in the context of adaptive Metropolis. It specifically **pinpoints two intuitively natural conditions** that quarantee convergence for adaptive algorithms:

Theorem. ([Roberts, Rosenthal, 2007]) *Assume that the adaptive MCMC chain* (X_n) *with kernels* P_C *satisfy*

1. *The sequence* $(M_\varepsilon(X_n, C_n))_{n=0}^{\infty}$ *is bounded in probability (***'containment'***) where* $M_\varepsilon(x,c) = \inf_n \{n \geq 1 : \|P_c^n(x,\cdot) - \pi\| \leq \varepsilon\}$,
2. $\lim_{n\to\infty} D_n = 0$, *in probability (***'diminishing adaptation '***), where* $D_n = \sup_x \|P_{C_n+1}(x,\cdot) - P_{C_n}(x,\cdot)\|$.

Then the algorithm is (weakly) ergodic.

The proof applies a nice coupling argument. Often it offers a lucid approach to convergence proofs, especially in some cases where the parameter C_n is **compactly confined** (i.e. the algorithm forces C_n to stay in a compact subset where e.g. a uniform geometric or polynomial ergodicity takes place for the kernels).

Generalizations

It appears that the 'containment' condition can be technically difficult to verify for many AM-type algorithms. Positive results exist, e.g.:

- Atchade and Fort prove convergence of adaptive algorithms in the subgeometric setup, assuming compact containment.
  Theorem. ([Atchade, Fort, 2010]) *The AM algorithm with (forced) compact containment converges assuming that* π *decays like* $|D^k log(\pi(x))| \sim |x|^{a-k}, \quad 0 \leq k \leq 2, \quad$ *for some* $a \in (0,1)$.
- Bai, Roberts, Rosenthal modify the AM algorithm by adding a fixed component in the transition:
  $\widetilde{P}_c = (1-\varepsilon)P_c + \varepsilon Q$ ('perform every now and then a move by a fixed proposal').
  Theorem. ([Bai, Roberts, Rosenthal, preprint 2008–2010]) *The AM algorithm* **with added fixed component** *converges if either*
  either (i) The conditions of the above theorem are valid (subgeometric case).
  or (ii) The condition (JH_1 *) is satisfied (geometric case).*
- [Andrieu, Thoms 2008] announces results by Andrieu and Tadic for some unconstrained algorithms (preprint containing full proof not yet published).

Convergence of the original AM

The first convergence result without modifications is contained in

- **Theorem.** ([Vihola, S., to appear in 2010]) *The original AM converges for targets satisfying (* JH_ρ *) for some* $\rho > 1$. *Also the central limit theorem is satisfied.*

  Structure of the proof in three steps:

  1. Prove a general convergence theorem for adaptive schemes with restricted growth rate for the parameter (e.g. force $|\widetilde{C}_n| \leq cn^\delta$).
  2. Prove that almost surely the original AM-chain satisfies the allowed growth rate.
  3. prove that AM-kernels satisfy the conditions of part 1.

  Proof of **1.** uses the approach of Andrieu and Moulines (Poisson kernels and the martingale decomposition), but requires even more careful quantitative estimates. No stochastic approximation techniques is used.
  Proof of **2.** not too difficult.
  Proof of **3.** is somewhat technical.

Not to forget the other talks!

Before mentioning open questions, let me point out that other interesting results will be given in the talks by e.g.

G. Fort

K. Latuszynski

M. Vihola

Open questions

We are not yet there ! Lot of **important, basic, questions remains open** in the theory of ergodicity of adaptive algorithms.

Here are some for the AM:

- How important is the ' ε ' in the definition of the AM-algorithm ?
  (Matti will tell us something interesting on this in his talk)
- Intuitively smoothness of the target is not at all necessary for ergodicity. Is this true?
- How slow can π decay?
  Is the existence of a finite second moment enough for basic ergodicity?

THANKS!

Literature mentioned in the talk:

- [Andrieu, Moulines, 2006] C. Andrieu, E.Moulines: *On the ergodicity properties of some adaptive MCMC algorithms,* Ann. Appl. Probab. 16 (2006), 1462–1505.
- [Andrieu, Robert, technical report 2001] C. Andrieu and C. P. Robert: *Controlled MCMC for optimal sampling*, Technical Report Ceremade 0125, Universite Paris Dauphine, 2001.
- [Andrieu, Thoms, 2008] C. Andrieu and J. Thoms: *A tutorial on adaptive MCMC'*, Stat Comput 18 (2008), 343–373.
- [Atchade, Fort, 1010] Y. Atchade and G. Fort: *Limit theorems for some adaptive MCMC algorithms with subgeometric kernels*, Bernoulli 16 (2010), 116–154.
- [Atchadé, Rosenthal, 2005] Y. F. Atchadé and J. S. Rosenthal: *On adaptive Markov chain Monte Carlo algorithms,* Bernoulli 11 (2005), 815–828.
- [Bai, Roberts, Rosenthal, Preprint 2008-2010] Y. Bai, G. O. Roberts, J. S. Rosenthal: *On the containment condition for adaptive Markov chain Monte Carlo algorithms*, preprint (2008–2010).
- [Gilks, Roberts, Sahu, 1998] W.R. Gilks, G. O. Roberts, S. K. Sahu: *Adaptive Markov chain Monte Carlo through regeneration,* J. Am. Stat. Assoc. 93 (1998), 1045–1054.
- [Haario, Saksman, Tamminen, 2001] H. Haario, E. Saksman, and J. Tamminen: *An adaptive Metropolis algorithm,* Bernoulli 7 (2001), 223–242.
- [Jarner, Hansen, 2000] S. F. Jarner and E. Hansen: *Geometric ergodicity of Metropolis algorithms,* Stochastic Process. Appl. 85 (2000), 341–361.
- [Roberts, Rosenthal, 2007] G. O. Roberts and J. S. Rosenthal: *Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms,* J. Appl. Probab. (2007), 458–475.
- [Saksman, Vihola, 2010] *On the ergodicity of the Adaptive Metropolis Algorithm in unbounded domains,* Ann. Appl. Probab. (2010), to appear.