On the ergodicity properties of some adaptive MCMC algorithms

C. Andrieu, E. Moulines

Arthur Gretton's notes

July 15, 2014

Why it's interesting

Metropolis Hastings samplers rely on having a good proposal distriubtion. How to get this?

1. Clever engineering?
2. HMC/MALA?
3. Adaptive proposal (simplest: Gaussian, more complex: mixture model)

Why it's interesting

Metropolis Hastings samplers rely on having a good proposal distriubtion. How to get this?

1. Clever engineering?
2. HMC/MALA?
3. Adaptive proposal (simplest: Gaussian, more complex: mixture model)

The challenge: how to adapt while preserving the desired target distribution?

Toy example: you can't just assume adaptation will work

Two state distribution $\mathsf{X} = \{1, 2\}$, $\theta \in \Theta := (0, 1)$,

$$P_\theta := \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}$$

stationary distribution is always $\pi := \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$.

Toy example: you can't just assume adaptation will work

Two state distribution $\mathsf{X} = \{1, 2\}$, $\theta \in \Theta := (0, 1)$,

$$P_\theta := \begin{bmatrix} 1-\theta & \theta \\ \theta & 1-\theta \end{bmatrix}$$

stationary distribution is always $\pi := \begin{pmatrix} 0.5 & 0.5 \end{pmatrix}$.
Now adaptive sampler: if in state 1, use $\theta(1)$, if in state 2, use $\theta(2)$,

$$P_\theta := \begin{bmatrix} 1-\theta(1) & \theta(1) \\ \theta(2) & 1-\theta(2) \end{bmatrix}.$$

Stationary distribution is now

$$\pi = \begin{bmatrix} \theta(2)/[\theta(1)+\theta(2)] & \theta(1)/[\theta(1)+\theta(2)] \end{bmatrix}.$$

Outline: adaptive M-H

Only for the adaptive Gaussian proposal (but can also work when the proposal is a mixture model)

1. Define the algorithm in two stages:
   1. A basic adaptive sampler that can get stuck.
   2. A more complex algorithm that re-initialises the simple algorithm when it gets stuck
2. When does adaptation work?
   1. Assuming the algorithm restarts only a finite number of times, what are the conditions for it to work?
   2. How do we know the algorithm will stop re-initialising? (the interesting bit)

Basics of M-H with Gaussian proposal

Metropolis Hastings. Given we are in state x,

- propse a candidate y using a proposal $q(y-x) = \mathcal{N}(x, \Gamma)$,
- accept with probability

$$\alpha(x,y) = \begin{cases} 1 \wedge \frac{\pi(y)}{\pi(x)} \frac{q(y-x)}{q(x-y)} & \text{if } \pi(x) q(x-y) > 0 \\ 1 & \text{otherwise} \end{cases}$$

- If we knew the covariance Γ_π of the target π, then there are heuristics for creating a good proposal.
- We do not know Γ_π, so we need to estimate proposal covariance from the sampler.

An adaptive sampler, Gaussian proposal

For case of a Gaussian proposal, Haario, Saksman, and Tamminen (2001) proposed the updates:

$$\mu_{k+1} = \mu_k + \gamma_{k+1}(X_{k+1} - \mu_k)$$
$$\Gamma_{k+1} = \Gamma_k + \gamma_{k+1}\left[(X_{k+1} - \mu_k)(X_{k+1} - \mu_k)^\top - \Gamma_k\right]$$

where γ_k are non-increasing positive stepsizes. A more concise but less clear notation:

$$\theta_{k+1} = \theta_k + \gamma_{k+1} H_{\theta_k}(X_{k+1})$$

where

$$\theta_k = [\mu_k \; \Gamma_k] \qquad H_\theta(X) = \left(x - \mu, (x - \mu)(x - \mu)^\top - \Gamma\right)^\top.$$

Does this work? If so, when?

Step 1: a basic adaptive sampler

A basic adaptive sampler uses:

1. A family P_θ of Markov transition kernels, where $P_\theta \pi = \pi \; \forall \, \theta \in \Theta$.
2. A family of update functions: $\{H_\theta(x) \; : \; \Theta \times \mathsf{X} \mapsto \Re^{n_\theta}\}$.
3. A "cemetary point" θ_c, where $\bar{\Theta} := \Theta \bigcup \{\theta_c\}$.
4. A sequence of stepsizes $\rho := \{\rho_k\}$ (non-increasing)

Run the sampler on the space (X_k, θ_k), with proposal density

$$Q_{\rho_k}(X_k, \theta_k; \underbrace{A \times B}_{\text{destination}}) = \int_A P_\theta(x, dy) \mathbb{I}\left\{\theta + \rho_k H(\theta_k, y) \in B\right\}$$

$$+ \delta_{\theta_c}(B) \int_A P_\theta(x, dy) \mathbb{I}\left\{\theta + \rho_k H(\theta_k, y) \notin \Theta\right\}$$

(where $B \in \mathcal{B}(\bar{\Theta})$). If the sampler gets in the cemetary state, $\theta_k = \theta_c$, then keep it there (it gets stuck).

Step 2: a sophisticated adaptive sampler with resets

For a more complex adaptive sampler (that does not get stuck), we need:

- A compact coverage $\{\mathcal{K}_q, q \geq 0\}$ of Θ ($\mathcal{K}_q$ are compact, Θ may be open):

$$\bigcup_{q\geq 0} \mathcal{K}_q = \Theta \quad \text{and} \quad \mathcal{K}_q \subset \operatorname{int}(\mathcal{K}_{q+1})$$

- A reset function:

$$\Pi \ : \ \mathsf{X} \times \bar{\Theta} \to \mathsf{K} \times \mathcal{K}_0$$

  where K a compact subset of X.
- A sequence of step sizes $\gamma := \{\gamma_k\}$ (non-increasing)

Step 2: a sophisticated adaptive sampler with resets

Define a Markov chain on

$$Z_k := \{X_k, \theta_k, \kappa_k, \nu_k\}$$

where:

- κ_k tells us which set $\mathcal{K}_{\kappa_k}$ we are in
- ν_k counts the number of samples since the last reset.

Step 2: a sophisticated adaptive sampler with resets

Define a Markov chain on

$$Z_k := \{X_k, \theta_k, \kappa_k, \nu_k\}$$

where:

- κ_k tells us which set $\mathcal{K}_{\kappa_k}$ we are in
- ν_k counts the number of samples since the last reset.

The **adaptive sampler** is defined as follows:

1. Draw $(X_k, \theta_k) \sim Q_{\gamma_{\kappa+\nu}}(X_k, \theta_k; \cdot)$ (the simple sampler)...
   1. *...unless* we have just reset ($\nu = 0$) in which case draw $(X_k, \theta_k) \sim Q_{\gamma_\kappa}(\Pi(X_k, \theta_k; \cdot))$.
2. When $\theta_k \in \mathcal{K}_\kappa$ set $\kappa_{k+1} = \kappa_k$ and $\nu_{k+1} = \nu_k + 1$,...
   1. ...*otherwise* reset the sampler: $\kappa_{k+1} = \kappa_k + 1$, $\nu_{k+1} = 0$.

Write as $\bar{\mathbb{E}}_\star, \bar{\mathbb{P}}_\star$ the expectations and probabilities under this chain.

Does the sampler work? Part 1

Assume that after a time, the sampler never resets again (which is the more interesting part to prove...):

$$\bar{\mathbb{P}}_\star \left(\sup_{n \geq 0} \kappa_n < \infty \right) = 1$$

What guarantee do we have?

Does the sampler work? Part 1

Assume that after a time, the sampler never resets again (which is the more interesting part to prove...):

$$\bar{\mathbb{P}}_\star \left(\sup_{n \geq 0} \kappa_n < \infty \right) = 1$$

What guarantee do we have?
The guarantee (and its conditions) are in terms of a norm: this norm is:

$$\|f\|_V = \sup_{x \in \mathsf{X}} \frac{|f(x)|}{V(x)} \qquad V : \mathsf{X} \to [1, \infty)$$

We will use

$$\|f\|_V = \sup_{x \in \mathsf{X}} \left(\frac{|f(x)|\, \pi(x)}{\sup_{x' \in \mathsf{X}} \pi(x')} \right) \qquad V(x) = \frac{\sup_{x' \in \mathsf{X}} \pi(x')}{\pi(x)}$$

Does the sampler work: Part 1

Assume the base sampler P_θ:

- **(A.1) converges fast** for any fixed $\theta \in \mathcal{K}$ in some *compact* $\mathcal{K}$. I.e. $\forall f \in \mathcal{L}_{V^r}$, $r \in [0,1]$, $\rho < 1$,

$$\left\| P_\theta^k f - \pi f \right\|_{V^r} \leq C \left\| f \right\|_{V^r} \rho^k$$

Does the sampler work: Part 1

Assume the base sampler P_θ:

- **(A.1) converges fast** for any fixed $\theta \in \mathcal{K}$ in some *compact* $\mathcal{K}$. I.e. $\forall f \in \mathcal{L}_{V^r}$, $r \in [0,1]$, $\rho < 1$,
$$\left\| P_\theta^k f - \pi f \right\|_{V^r} \leq C \left\| f \right\|_{V^r} \rho^k$$
- **(A.2) does not change much** when θ changes within $\mathcal{K}$: for $\theta, \theta' \in \mathcal{K}$,
$$\left\| P_\theta f - P_{\theta'} f \right\|_{V^r} \leq C \left\| f \right\|_{V^r} \left| \theta - \theta' \right|$$

Does the sampler work: Part 1

Assume the base sampler P_θ:

- **(A.1) converges fast** for any fixed $\theta \in \mathcal{K}$ in some *compact* $\mathcal{K}$. I.e. $\forall f \in \mathcal{L}_{V^r}$, $r \in [0,1]$, $\rho < 1$,

$$\left\| P_\theta^k f - \pi f \right\|_{V^r} \leq C \left\| f \right\|_{V^r} \rho^k$$

- **(A.2) does not change much** when θ changes within $\mathcal{K}$: for $\theta, \theta' \in \mathcal{K}$,

$$\left\| P_\theta f - P_{\theta'} f \right\|_{V^r} \leq C \left\| f \right\|_{V^r} \left| \theta - \theta' \right|$$

Assume the adaptive mapping H_θ is **well behaved (A.3)**: $\beta \in [0, 1/2]$,

$$\sup_{(\theta,\theta') \in \mathcal{K}\times\mathcal{K}, \theta \neq \theta'} \left| \theta' - \theta \right|^{-1} \left\| H_\theta - H_{\theta'} \right\|_{V^\beta} < \infty.$$

Does the sampler work: Part 1

Assume the base sampler P_θ:

- **(A.1) converges fast** for any fixed $\theta \in \mathcal{K}$ in some *compact* $\mathcal{K}$. I.e. $\forall f \in \mathcal{L}_{V^r}$, $r \in [0,1]$, $\rho < 1$,

$$\left\| P_\theta^k f - \pi f \right\|_{V^r} \leq C \left\| f \right\|_{V^r} \rho^k$$

- **(A.2) does not change much** when θ changes within $\mathcal{K}$: for $\theta, \theta' \in \mathcal{K}$,

$$\left\| P_\theta f - P_{\theta'} f \right\|_{V^r} \leq C \left\| f \right\|_{V^r} \left| \theta - \theta' \right|$$

Assume the adaptive mapping H_θ is **well behaved (A.3)**: $\beta \in [0, 1/2]$,

$$\sup_{(\theta,\theta') \in \mathcal{K}\times\mathcal{K}, \theta \neq \theta'} \left| \theta' - \theta \right|^{-1} \left\| H_\theta - H_{\theta'} \right\|_{V^\beta} < \infty.$$

Then **(Theorem 8)** as long as $\sum_{k=1}^{\infty} k^{-1}\gamma_k < \infty$,

$$n^{-1} \sum_{k=1}^{n} \left[f(X_k) - \pi(f) \right] \xrightarrow{\text{a.s.}}_{\bar{\mathbb{P}}_\star} 0.$$

Does the sampler work: Part 2

Why does the sampler stop resetting?

Consider the optimization:

$$\theta_{k+1} = \theta_k + \gamma_{k+1} h(\theta_k) + \gamma_{k+1}\xi_{k+1}$$

where:

$$\begin{aligned} X_{k+1} &\sim P_{\theta_k}(X_k, \cdot) \\ h(\theta) &= \int_{\mathsf{X}} H(\theta, x)\pi(dx) \\ \xi_k &= H(\theta_{k-1}, X_k) - h(\theta_{k-1}) \end{aligned}$$

We want this to converge to the set $\theta \in \Theta, h(\theta) = 0$. This is a ***stochastic optimization*** problem.

Does the sampler work: Part 2

To prove the stability of this sampler: define a Lyapunov function $w : \Theta \to [0, \infty)$, where

$$\langle \nabla w(\theta), h(\theta) \rangle \leq 0.$$

The set of stationary points of the optimization is written

$$\mathcal{Z} := \{\theta \in \Theta \,:\, \langle \nabla w(\theta), h(\theta) \rangle = 0\}.$$

Under some technical conditions on w **(A4)**:

1. $\mathcal{W}_M := \{\theta \in \Theta, w(\theta) \leq M\}$ is compact $\forall M > 0$
2. $\mathcal{Z} \in \mathrm{int}(\Theta)$
3. The closure of $w(\mathcal{Z})$ has an empty interior

and on the stepsizes **(A5)**:

$$\sum_{k=1}^{\infty} \gamma_k = \infty \qquad \sum_{k=1}^{\infty} \left\{\gamma_k^2 + k^{-1/2}\gamma_k\right\} < \infty,$$

then **(Theorem 11)** the number of resets is a.s. finite, and θ_k converges to a point in $\mathcal{L}$.

Illustration: Gaussian case

In the case of the Gaussian sampler of Haario, Saksman, and Tamminen (2001), the Lyapunov function is

$$w(\mu, \Gamma) = \log\det\Gamma + (\mu - \mu_\pi)^\top \Gamma^{-1}(\mu - \mu_\pi) + \mathrm{Tr}(\Gamma^{-1}\Gamma_\pi).$$

The set $\mathcal{Z} := \{\theta \in \Theta \,:\, \langle \nabla w(\theta), h(\theta)\rangle = 0\}$ contains a single point **(Lemma 14)**:

$$\mathcal{L} := \{\mu_\pi, \Gamma_\pi\}.$$

Illustration: Gaussian case

In the case of the Gaussian sampler of Haario, Saksman, and Tamminen (2001), the Lyapunov function is

$$w(\mu, \Gamma) = \log\det\Gamma + (\mu - \mu_\pi)^\top \Gamma^{-1}(\mu - \mu_\pi) + \mathrm{Tr}(\Gamma^{-1}\Gamma_\pi).$$

The set $\mathcal{Z} := \{\theta \in \Theta \,:\, \langle \nabla w(\theta), h(\theta)\rangle = 0\}$ contains a single point **(Lemma 14)**:

$$\mathcal{L} := \{\mu_\pi, \Gamma_\pi\}.$$

The sampler (with resets!) is guaranteed to converge **(Theorem 15)**: for any $f \in \mathcal{L}(w^\alpha)$,

$$n^{-1}\sum_{k=1}^{n}\left(F(X_k) - \int_{\mathsf{X}} f(x)\pi(x)dx\right) \xrightarrow{\text{a.s.}}_{\bar{\mathbb{P}}_*} 0.$$

Illustration: Gaussian case

In the case of the Gaussian sampler of Haario, Saksman, and Tamminen (2001), the Lyapunov function is

$$w(\mu, \Gamma) = \log\det\Gamma + (\mu - \mu_\pi)^\top \Gamma^{-1}(\mu - \mu_\pi) + \mathrm{Tr}(\Gamma^{-1}\Gamma_\pi).$$

The set $\mathcal{Z} := \{\theta \in \Theta \ : \ \langle \nabla w(\theta), h(\theta) \rangle = 0\}$ contains a single point **(Lemma 14)**:

$$\mathcal{L} := \{\mu_\pi, \Gamma_\pi\}.$$

The sampler (with resets!) is guaranteed to converge **(Theorem 15)**: for any $f \in \mathcal{L}(w^\alpha)$,

$$n^{-1} \sum_{k=1}^{n} \left(F(X_k) - \int_{\mathsf{X}} f(x)\pi(x)dx \right) \overset{\text{a.s.}}{\longrightarrow}_{\bar{\mathbb{P}}_*} 0.$$

The analsys can also be done for a mixture model proposal fit by an online EM algorithm **(Section 7)**.