Adaptive MCMC

5th July 2023

Table of Contents

Table of Contents

MCMC Overview

Goal: obtain a Markov chain $X_1, X_2, \ldots$ with transition P on $\mathcal{X} \subset \mathbb{R}^d$ that has stationary distribution π ("π-ergodicity"[1]). Then we can approximate π-integrable functions

$$I(f) = \int_{\mathcal{X}} f(x)\pi(dx)$$

by

$$\hat{I}_N(f) := \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

(though perhaps ignoring the first few samples $X_1, \ldots, X_{i_0}$ for some $i_0 \in \mathbb{N}$ as *burn-in* to allow the chain to mix sufficiently and reach the distribution π).

[1]Defined broadly in Andrieu and Thoms 2008

MCMC Overview

Metropolis-Hastings[2] (MH) at each step $i = 0, 1, \ldots$:

1. Propose $Y_{i+1} \sim q(X_i, \cdot)$
2. Set $X_{i+1} = Y_{i+1}$ with probability

$$\alpha(X_i, Y_{i+1}) = \min\left(1, \frac{\pi(Y_{i+1})q(Y_{i+1}, X_i)}{\pi(X_i)q(X_i, Y_{i+1})}\right),$$

   otherwise $X_{i+1} = X_i$.

E.g. Normal Symmetric Random Walk Metropolis (N-SRWM):

$$q_\theta(X_i, Y_{i+1}) = \mathcal{N}(Y_{i+1}; X_i, \theta^2 I_d)$$

for some $\theta > 0$. The corresponding estimator $\hat{I}_N^\theta(f)$ has high variance for values of θ that are too small or too large (the same can happen with non-isotropic proposal covariances in place of $\theta^2 I_d$).

[2] Metropolis et al. 1953; Hastings 1970.

Table of Contents

Adaptive MCMC Overview

Some theoretical results exist for the optimal proposals in different scenarios:

- e.g. using a multivariate random walk

$$Y_{i+1} \sim \mathcal{N}(X_i, 2.38^2 C/d)$$

  where d is the dimension of $\mathcal{X}$ and C is the covariance of the target distribution π, which is a mixture of Gaussians (or just has a large dimension d)[3].

But Adaptive MCMC algorithms aim to find such a θ automatically in a wider setting.

The general adaptive MCMC game:

- Given some set of proposal parameters Θ.
- Choose some $\theta_i \in \Theta$ at each step i (given $X_0, \ldots, X_{i-1}, Y_1, \ldots, Y_{i-1}$ and θ_{i-1}) and use transition P_{θ_i} to generate X_{i+1}.
- Eventually we want to stop adapting and use the same θ for all steps (at least with high probability).

[3] Gareth O. Roberts and Jeffrey S. Rosenthal 2001.

Table of Contents

Ensuring π-ergodicity

In order to achieve π-ergodicity of our adaptive process, so that

$$|\mathbb{E}(f(X_i)) - \mathbb{E}_\pi(f(X))| \to 0$$

as $i \to \infty$ for any $f : \mathcal{X} \to \mathbb{R}$, we require[4]:

1. Stationarity: Every $\theta \in \Theta$ has π-ergodicity.
2. Diminishing Adaptation: The 'amount' of adaptation decreases as $i \to \infty$,

$$\lim_{i\to\infty} \sup_{X\in\mathcal{X}} ||P_{\theta_{i+1}}(X,\cdot) - P_{\theta_i}(X,\cdot)|| = 0$$

   (in probability). This is usually achieved by making sure adaptations:
   - are small with high probability, or
   - take place with probability $p(i) \to 0$ as $i \to \infty$ (e.g. stop adapting after τ steps).
3. Containment: Times from X_i to stationary distribution π are bounded in probability as $i \to \infty$. (This is usually achieved as a result of the two conditions above, depending on how diminishing adaptation is achieved.)

[4] Gareth O Roberts and Jeffrey S Rosenthal 2005.

WLLN (for bounded functions)

Under stationarity, adaptation and containment we get:

$$\frac{\lim_{n\to\infty}\sum_{i=1}^{n} f(X_i)}{n} = \pi(f)$$

in probability for any bounded function f.

(But, convergence for all $\mathbf{L^1}$ functions does not follow[5].)

[5] Yang 2008.

When containment fails

Containment fails when different subsets $\mathcal{K} \subset \Theta$ of parameters converge to π in 'different ways'—without a "common drift function".

Solution: Limit the subset of Θ that is explored during adaptation in order to avoid the "bad" values for which convergence to π can take arbitrarily long (often at the boundary of Θ).

- Truncate Θ to exclude these "bad" values.
  - Requires some knowledge of the problem at hand, but sometimes this can be found by considering a desired drift function (e.g. G. O. Roberts and Tweedie 1996; Atchadé 2006).
- Andrieu and Thoms 2008—"vanishing adaptation" (i.e. no adaptation after a certain step $\tau \in \mathbb{N}$) is sufficient for containment.

Convergence towards π

Assume we have a subset of "good" values $\mathcal{K} \subset \Theta$ for which containment is ensured (i.e. for which there is a common drift function), and let σ be the first time i at which $\theta_i \notin \mathcal{K}$ (this may be infinity).
Then under certain conditions[6] (satisfied by N-SRWM), with "smoothly decaying" step-sizes $|\theta_i - \theta_{i-1}| \leq \gamma_i$ (e.g. $\gamma_i = i^{-\alpha}, \alpha > 0$), there exists a constant $C' > 0$ such that for all $i \geq 1$ and $|f| \leq 1$:

$$|\mathbb{E}[(f(X_i) - \mathbb{E}_\pi(f)) \underbrace{\mathbb{I}\{\sigma \geq i\}}_{\substack{\text{only consider} \\ \theta_i \in \mathcal{K}}}]| < C'\gamma_i.$$

That is, whilst θ doesn't leave $\mathcal{K}$, convergence to π occurs at a rate of at least $\{\gamma_i\}$—and doesn't not require convergence of $\{\theta_i\}$!

[6] Andrieu and Moulines 2006.

Monte Carlo Error

Bias for a single sample X_i:

$$|\mathbb{E}[(f(X_i) - \mathbb{E}_\pi(f)) \underbrace{\mathbb{I}\{\sigma \geq i\}}_{\substack{\text{only consider} \\ \theta_i \in \mathcal{K}}}]| < C'\gamma_i.$$

It can then be proved that there exist constants $A(\gamma, \mathcal{K})$ and $B(\gamma, \mathcal{K})$ such that for any $n \geq 1$ the error is bounded as:

$$\sqrt{\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\pi(f))\right|^2 \mathbb{I}\{\sigma \geq i\}\right]} \leq \underbrace{\frac{A(\gamma, \mathcal{K})}{\sqrt{n}}}_{\text{standard Monte Carlo error}} + \underbrace{B(\gamma, \mathcal{K})\frac{\sum_{i=1}^{n}\gamma_i}{n}}_{\text{price paid for adaptation}}$$

(So if $\gamma_i = i^{-\alpha}$, $\alpha \in (0, 1)$, then $\frac{\sum_{i=1}^{n}\gamma_i}{n} \sim \frac{N^{-\alpha}}{1-\alpha}$, meaning there is no loss in rate of convergence for $\alpha \geq 1/2$.)

Table of Contents

Adaptive MCMC Algorithms

- Random Walk Metropolis (RWM):

$$q(X_i, dX) = \mathcal{N}(X_i, s_d \boldsymbol{\Sigma})$$

  for matrix $\boldsymbol{\Sigma}$ and scaling factor $s_d > 0$.

  - Very popular and fairly simple
  - Many variations: component-wise, Metropolis-within-Gibbs (MwG), PCA-based, &c.
  - Lots of theoretical results

  We'll start with Haario et al.'s "Adaptive Metropolis"[7] (AM) and then look at variations.

- Metropolis-Adjusted Langevin Algorithm (MALA)[8]: for a matrix $\boldsymbol{\Sigma}$,

$$q_\theta(X_i, dX) = \mathcal{N}(X_i + \boldsymbol{\Sigma} \nabla \log \pi(X)/2, \boldsymbol{\Sigma}).$$

  Can have faster convergence for high-dimensional proposals than RWM.

[7] Haario, Saksman, and Tamminen 2001.

[8] Walter R. Gilks, Gareth O. Roberts, and Sahu 1998.

RWM: $q(X, dX) = \mathcal{N}(X, s_d\boldsymbol{\Sigma})$

Theoretical result[9]: for a wide range of target distributions, optimal proposal for RWM is with $\boldsymbol{\Sigma} = C$ and $s_d = 2.38^2/d$ where d is the dimension of $\mathcal{X}$ and C is the covariance of π.

Haario et al.'s "Adaptive Metropolis"[10] (AM) uses this result to adapt $\boldsymbol{\Sigma}$ at each step i, using an empirical covariance $\hat{C}_i$ multiplied by $s_d = 2.38^2/d$.

In general, begin with some initial $\hat{C}_0$ and $i_0 \in \mathbb{N}$ initial steps without adaptation.

$$\hat{C}_i = \begin{cases} \hat{C}_0 & i \leq i_0 \\ s_d \mathrm{cov}(X_0, \ldots, X_{i-1}) + s_d \epsilon I_d & i > i_0 \end{cases}$$

where $s_d > 0$ is a scale factor, $\epsilon > 0$ is a small constant (used to avoid singularity of $\hat{C}_i$—particularly in multimodal posteriors—and required for Haario's proof of AM's π-ergodicity).

[9] Gelman, G. O. Roberts, and W. R. Gilks 1996.

[10] Haario, Saksman, and Tamminen 2001.

AM: Efficient Updates

Using the fact that

$$\mathrm{cov}(X_0, ..., X_i) = \frac{1}{i}\left(\sum_{k=0}^{i} X_k^T X_k - (i+1)\bar{X}_i\bar{X}_k^T\right),$$

where $\bar{X}_i = \frac{1}{i}\sum_{k=0}^{i} X_k$, we can update $\hat{C}_i$ incrementally[11]:

$$\hat{C}_{i+1} = \frac{i-1}{i}\hat{C}_i + \frac{s_d}{i}(i\bar{X}_{i-1}\bar{X}_{i-1}^T - (i+1)\bar{X}_i\bar{X}_{i-1}^T + X_i X_i^T + \epsilon I_d).$$

[11](I *think* that this is essentially the same as the "Rao-Blackwellised AM algorithm" presented by Andrieu and Thoms 2008.)

AM: Adapting the scale factor s_d

Using $s_d = 2.38^2/d$ isn't always optimal (e.g. for multimodal non-Gaussian-mixture posteriors), so we can adapt s_d too.

The other common type of theoretical result is the optimal acceptance rate α^* for a given proposal and target distribution family:

- For full-rank multivariate Gaussian proposals, $\alpha^* = 0.234$.
- For individual components of a multivariate Gaussian proposal, $\alpha^* = 0.44$
  - (often here the optimal proposal is $\mathcal{N}(X_i^{(j)}, 2.4^2\xi_i^{(j)})$ where $\xi_i^{(j)}$ is the target *conditional* variance of the jth component).

Adapting s_d is particularly useful at the start of the algorithm, when our covariance estimate is likely to be poor.

Then we can use Robbins-Monro style updates to optimise $\theta = s_d$ such that $\alpha_i(\theta) \to \alpha^*$ as $i \to \infty$.

AM: Optimising s_d via Robbins-Monro

We want to match a target acceptance rate α^*:

1. One-dimensional updates: $\alpha^* = 0.44$.
2. Multivariate updates: $\alpha^* = 0.234$.

Robbins-Monro updates: with $\theta = s_d$ and non-negative step sizes $\{\gamma_i\}$,

$$\theta_{i+1} = \theta_i - \gamma_i(\bar{\alpha}_i(\theta) - \alpha^*),$$

where $L \in \mathbb{N}$, $Y_{i,1}, ..., Y_{i,L} \sim q_\theta(X_i, \cdot)$ are IID and

$$\bar{\alpha}_i(\theta) = \frac{1}{L}\sum_{l=1}^{L} \min\left(1, \frac{\pi(Y_{i,l})q_\theta(Y_{i,l}, X_i)}{\pi(X)q_\theta(X_i, Y_{i,l})}\right).$$

Intuition:

- if $\bar{\alpha}_i(\theta)$ is too high ($\bar{\alpha}_i(\theta) - \alpha^* > 0$), make proposal tighter by reducing $\theta = s_d$,
- if $\bar{\alpha}_i(\theta)$ is too low ($\bar{\alpha}_i(\theta) - \alpha^* < 0$), make proposal wider by increasing $\theta = s_d$.

AM: Generic Robbins-Monro Updates

Generic Robbins-Monro updates for any suitable parameterisation θ of the proposal q_θ:

$$\theta_{i+1} = \theta_i - \gamma_i H(\theta_i, X_0, \ldots, Y_i, X_i, Y_{i+1}, X_{i+1})$$

for some $H : \Theta \times \mathcal{X}^{1+2(i+1)} \to \Theta$ (note we have access to discarded proposals Y_k).

This is to find roots of the equation $H(\theta) = 0$.

(In the previous slide, $\Theta = \mathbb{R}^+$ and $H(\theta_i, X_0, \ldots, Y_i, X_i, Y_{i+1}, X_{i+1}) = \bar{\alpha}_i(\theta) - \alpha^*$.)

AM: Moment Matching

$$\theta_{i+1} = \theta_i - \gamma_i H(\theta_i, X_0, \ldots, Y_i, X_i, Y_{i+1}, X_{i+1})$$

Moment matching: With $\mu_\pi, \mathbf{\Sigma}_\pi$ the true mean and covariance of π and $\mu(\theta), \mathbf{\Sigma}(\theta)$ are the empirical mean and covariance, try to find θ for which

$$(\mu_\pi, \mathbf{\Sigma}_\pi) = (\mu(\theta), \mathbf{\Sigma}(\theta))$$

Under certain conditions, this can be shown[12] to be equivalent to minimising the KL, in which case we end up with

$$H(X, \theta) = \nabla_\theta \log \frac{\pi(X)}{q_\theta(X)}$$

[12] Andrieu and Moulines 2006.

AM: VI Updates

$$\theta_{i+1} = \theta_i - \gamma_i H(\theta_i, X_0, \ldots, Y_i, X_i, Y_{i+1}, X_{i+1})$$

$$H(X, \theta) = \nabla_\theta \log \frac{\pi(X)}{q_\theta(X)}$$

- This is just VI with a Gaussian approximate posterior (and with a Metropolis acceptance step).
- Not sure this is very promising: no guarantee $\exists \theta$ s.t. $q_\theta = \pi$.
- But, we could use several separate (Gaussian) proposals for different parts of π (e.g. for each latent r.v.) and tune these each with VI (with optional covariance scaling factors).

AM: A stopping rule

Stop adaptation once we see that

$$\frac{1}{n}\sum_{i=1}^{n} H(\theta_i, X_{i+1})$$

stabilises and does not change by more than some small $\varepsilon > 0$ for $m \in \mathbb{N}$ consecutive iterations.

"More principled statistical rules relying on the CLT can also be suggested, but we do not expand on this here"[13].

[13] Andrieu and Thoms 2008.

AM: Adaptive step size

Schemes for step sizes $\{\gamma_i\}$:

1. Deterministic and non-increasing e.g. $\gamma_i = i^{-\alpha}$, $\alpha > 0$.
2. Random with $\gamma_i \in \{0, \delta\}$ such that $\mathbb{P}(\gamma_i = \delta) = p_i$, where $\{p_i\}$ deterministic and non-increasing s.t. $p_i \to 0$ as $i \to \infty$. But "it is not always clear what the advantage of introducing such an additional level of randomness is"[14].
3. Various automatic choices based on θ_i and X_i given a predefined function $\gamma : [0, \infty) \to [0, \infty)$. Typically based on the idea that alternating signs of $H(\theta, X)$ tend to suggest θ_i is oscillating around a solution. E.g.:
   - With $\langle u, v \rangle$ denoting the inner product between vectors u and v,

     $$\gamma_i = \gamma\left(\sum_{k=1}^{i-1} \mathbb{I}\{\langle H(\theta_{k-1}, X_k), H(\theta_k, X_{k+1})\rangle \leq 0\}\right).$$

   - Same as above[15] but with separately derived step sizes for each component of θ.

[14] Andrieu and Thoms 2008.

[15] Delyon and Juditsky 1993.

AM: Other Variations

- Metropolis-within-Gibbs (MwG) with multivariate proposals that *aren't* full rank in terms of $dim(\mathcal{X}) = d$.
- Update in the direction of a sampled principal component (with more important PCs more likely to be sampled) using online PCA.
  - (Distance along this direction is sampled from a RWM proposal)[16].
- Online EM algorithm version that uses Gaussian mixture proposals[17].

[16] Andrieu and Thoms 2008.
[17] Andrieu and Moulines 2006.

Metropolis-Adjusted Langevin Algorithm (MALA)

Perform AM as before (and all the variations that we've covered), but with a Langevin proposal (thus using drift function $\nabla \log \pi(X)$):

$$q_\theta(X, dX) = \mathcal{N}(X + \boldsymbol{\Sigma}\nabla \log \pi(X)/2, \boldsymbol{\Sigma}).$$

- Typically still use $\boldsymbol{\Sigma} = s_d C$ for some scaling factor $s_d > 0$ and covariance C of π (or an estimate thereof).
- Optimal acceptance rate is typically $\alpha^* = 0.574$ in most situations.

Popular variation: Truncated drift MALA (T-MALA)[18]—solves some of MALA's convergence problems by truncating the drift function to avoid "bad" values of θ.

$$\nabla \log \pi(X) \mapsto \frac{\delta}{\max(\delta, |\nabla \log \pi(X)|)} \nabla \log \pi(X)$$

where $\delta > 0$.

[18] Atchadé 2006.

A Quick Comparison of the Methods

Generally speaking...

- MALA has fastest convergence for multivariate proposals
  - (Optimal convergence time is $\mathcal{O}(d^{1/3})$ compared to $\mathcal{O}(d)$ for RWM),
- But MALA is less robust to light tails, discontinuous densities and very sub-optimal for single-component updates.
  - (Although T-MALA aims to solve some of these problems).
- RWM is very robust to a wide variety of distributions, with component-wise versions/Metropolis-within-Gibbs being at least as good (when sensibly scaled).
- Full multivariate RWM tends to converge to the same proposals as component-wise/MwG proposals, but often more slowly.

Table of Contents

Massively Parallel MCMC

In massively parallel MCMC, at each iteration we have indexed latent samples $z^{\mathbf{k}} \in \mathcal{Z}$ (where $\mathbf{k} = (k_1, ..., k_n) \in \{1, ..., K\}^n$ is a tuple of indices for of our n latent variables) and we want to generate new 'unindexed' samples $z^{/\mathbf{k}} \in \mathcal{Z}^{K-1}$.
The proposals that we use for the jth latent variable must be

- independent of all other variables,

$$q(z_j^{/k_j}; x, z^{\mathbf{k}}, z_{\mathrm{qa}(j)}^{/\mathbf{k}}) = q(z_j^{/k_j}; z_j^{k_j}),$$

- symmetric w.r.t. the choice of k_j, in the sense that for any $k_j' \neq k_j$,

$$q(z_j^{/k_j}; z_j^{k_j}) = q(z_j^{/k_j'}; z_j^{k_j'}).$$

RWM satisfies these, as does (T-)MALA, so we should be able to use the adaptive schemes discussed above.

Massively Parallel Adaptive MCMC

Recall the two main adaptive strategies (leading to functions H):

1. Try to reach a target acceptance rate α^* by adapting s_d in the AM algorithm.
2. Moment matching/VI with a Metropolis acceptance step[19].

In the massively parallel setting, we can do the following *very* fast:

1. Compute moments—useful for AM algorithm.
   - (Including with the AMMP-IS moving average thing over MH iterations?)
2. Perform VI.

So both adaptive schemes seem promising (and hopefully not too complicated), both with RWM and (T-)MALA proposals.

---

[19] There are a few more details involved/variations possible in this.

Conclusion

- Adaptive MCMC is a *very* big field with an endless number of variations for each algorithm.
- But in general it seems that RWM and MALA are the most popular proposal types.
- In particular, the basic AM algorithm (and its variations) seems like a good starting point for massively parallel adaptive MCMC.

References I

Andrieu, Christophe and Éric Moulines (Aug. 2006). "On the ergodicity properties of some adaptive MCMC algorithms". In: *The Annals of Applied Probability* 16.3. Publisher: Institute of Mathematical Statistics, pp. 1462–1505. ISSN: 1050-5164, 2168-8737. DOI: 10.1214/105051606000000286. URL: https://projecteuclid.org/journals/annals-of-applied-probability/volume-16/issue-3/On-the-ergodicity-properties-of-some-adaptive-MCMC-algorithms/10.1214/105051606000000286.full (visited on 06/15/2023).

Andrieu, Christophe and Johannes Thoms (Dec. 2008). "A tutorial on adaptive MCMC". en. In: *Statistics and Computing* 18.4, pp. 343–373. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-008-9110-y. URL: http://link.springer.com/10.1007/s11222-008-9110-y (visited on 06/15/2023).

References II

Atchadé, Yves F. (June 2006). "An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift". en. In: *Methodology and Computing in Applied Probability* 8.2, pp. 235–254. ISSN: 1387-5841, 1573-7713. DOI: 10.1007/s11009-006-8550-0. URL: http://link.springer.com/10.1007/s11009-006-8550-0 (visited on 06/15/2023).

Delyon, Bernard and Anatoli Juditsky (Nov. 1993). "Accelerated Stochastic Approximation". In: *SIAM Journal on Optimization* 3.4. Publisher: Society for Industrial and Applied Mathematics, pp. 868–881. ISSN: 1052-6234. DOI: 10.1137/0803045. URL: https://epubs.siam.org/doi/10.1137/0803045 (visited on 07/04/2023).

Gelman, A., G. O. Roberts, and W. R. Gilks (1996). "Efficient Metropolis jumping rules". In: *Bayesian Statistics*. Ed. by J. M. Bernardo et al. Oxford University Press, Oxford, pp. 599–608.

References III

Gilks, Walter R., Gareth O. Roberts, and Sujit K. Sahu (1998). “Adaptive Markov Chain Monte Carlo through Regeneration”. In: *Journal of the American Statistical Association* 93.443. Publisher: [American Statistical Association, Taylor & Francis, Ltd.], pp. 1045–1054. ISSN: 0162-1459. DOI: 10.2307/2669848. URL: https://www.jstor.org/stable/2669848 (visited on 06/15/2023).

Haario, Heikki, Eero Saksman, and Johanna Tamminen (2001). “An Adaptive Metropolis Algorithm”. In: *Bernoulli* 7.2. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability, pp. 223–242. ISSN: 1350-7265. DOI: 10.2307/3318737. URL: https://www.jstor.org/stable/3318737 (visited on 06/15/2023).

Hastings, W K (1970). “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”. en. In.

References IV

- Metropolis, Nicholas et al. (June 1953). "Equation of State Calculations by Fast Computing Machines". en. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.1699114. URL: http://aip.scitation.org/doi/10.1063/1.1699114 (visited on 02/03/2023).
- Roberts, G. O. and R. L. Tweedie (Mar. 1996). "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms". In: *Biometrika* 83.1, pp. 95–110. ISSN: 0006-3444. DOI: 10.1093/biomet/83.1.95. URL: https://doi.org/10.1093/biomet/83.1.95 (visited on 07/03/2023).
- Roberts, Gareth O and Jeffrey S Rosenthal (Mar. 2005). "Coupling and Ergodicity of Adaptive MCMC". en. In.
- Roberts, Gareth O. and Jeffrey S. Rosenthal (2001). "Optimal Scaling for Various Metropolis-Hastings Algorithms". In: *Statistical Science* 16.4. Publisher: Institute of Mathematical Statistics, pp. 351–367. ISSN: 0883-4237. URL: https://www.jstor.org/stable/3182776 (visited on 07/03/2023).

References V

Yang, Chao (2008). “Ergodicity of Adaptive MCMC and its Applications”. en. PhD thesis. University of Toronto. URL: http://probability.ca/jeff/ftpdir/chaothesis.pdf.