Error bounds for adaptive MCMC within doubly intractable distributions

Algorithms and Computationally Intensive Inference seminar – Warwick

Julian Hofstadler – University of Passau

ongoing work with

B. Eltzner, B. de Groot, M. Habeck and D. Rudolf

October 20, 2023

CRC 1456
MATHEMATICS OF
EXPERIMENT

Setting

- Observation space $\mathcal{Y}$, equipped with some reference measure μ.
- Parameter space Θ, equipped with some probability measure ν.
- A non-negative function $\varrho\colon \mathcal{Y} \times \Theta \to [0, \infty)$ such that

$$\forall \theta \in \Theta: \qquad Z(\theta) = \int_{\mathcal{Y}} \varrho(y|\theta)\mu(\mathrm{d}y) \in (0, \infty).$$

Setting – II

Given $y \in \mathcal{Y}$ consider the density (w.r.t. (Θ, ν)):

$$p(\theta|y) = \frac{1}{C_y} \frac{\varrho(y|\theta)}{Z(\theta)},$$

with normalizing constant $C_y = \int_\Theta \frac{\varrho(y|\theta)}{Z(\theta)} \, \nu(\mathrm{d}\theta)$.

One can interpret $p(\theta|y)$ as posterior density with

- Likelihood function $\frac{\varrho(y|\theta)}{Z(\theta)}$.
- Prior distribution ν.

Setting – III

Goal: Given observation $\bar{y}$ extract information from distribution π, where

$$\pi(\mathrm{d}\theta) = p(\theta|\bar{y})\nu(\mathrm{d}\theta) = \frac{1}{C_{\bar{y}}} \frac{\varrho(\bar{y}|\theta)}{Z(\theta)} \nu(\mathrm{d}\theta).$$

Problems:

- Evaluating $Z(\cdot)$ is computationally infeasible.
- Normalizing constant $C_{\bar{y}}$ is not known.

So π contains two unknown quantities and is therefore called doubly intractable.

However, we still want to extract information from π.

Classical MCMC in our setting

Classical MCMC methods use a Markov Chain $(\widetilde{\xi}_n)_{n\in\mathbb{N}_0}$ such that:

- $(\widetilde{\xi}_n)_{n\in\mathbb{N}_0}$ has kernel K, i.e., $\mathbb{P}(\widetilde{\xi}_n \in A | \widetilde{\xi}_{n-1} = \theta_{n-1}) = K(\theta_{n-1}, A)$.
- The 'limit distribution' is given by π.
- Limit theorems ensure that $(\widetilde{\xi}_1, \ldots, \widetilde{\xi}_n)$ approximates π (in some sense).

What is required for implementing such a method?

- Not knowing $C_{\bar{y}}$ is fine.
- Not having access to $Z(\theta)$ is problematic.

Methods for the doubly intractable setting

A number of methods have been suggested and studied to tackle the doubly intractable setting, including

- Approximate Bayesian Computation (ABC) algorithms Marin et al. (2012); ...
- Noisy MCMC Alquier et al. (2016); Habeck et al. (2020); ...
- Pseudo-marginal methods Andrieu and Roberts (2009); ...
- Auxiliary variable methods Møller et al. (2006); Murray et al. (2006); ...
- ...

Let me also recommend the review paper Park and Haran (2018) about MCMC methods in the doubly intractable setting.

Our approach: Adaptive MCMC

Recently, there were promising results in biophysics Eltzner et al. (2023), Habeck (2014).

Additionally, one can prove that adaptive MCMC methods work in the doubly intractable setting Atchadé et al. (2013); Liang et al. (2016).

Goal: Better understanding of error behaviour of adaptive MCMC.

This talk:

- Introduce (some basics) of adaptive MCMC.
- Provide error bounds in the doubly intractable setting.
- Have a look at a toy example, the Ising model.

Some basics of adaptive MCMC

Adaptive MCMC methods construct a sequence of random variables $(\widetilde{\xi}_n)_{n\in\mathbb{N}_0}$ via

- A family of transition kernels $\{K_\gamma\}_\gamma$.
- A sequence of random variables $(\Gamma_n)_{n\in\mathbb{N}}$.

Each $\widetilde{\xi}_n$ is specified by

$$\mathbb{P}\left[\widetilde{\xi}_n \in A \mid \widetilde{\xi}_{n-1} = x, \Gamma_{n-1} = \gamma\right] = K_\gamma(x, A).$$

- $(\widetilde{\xi}_n)_{n\in\mathbb{N}}$ in general is non-markovian.
- Each transition kernel K_γ may have a different invariant distribution π_γ.

Back to the doubly intractable setting

Idea: Approximate Z by Z_n and work with

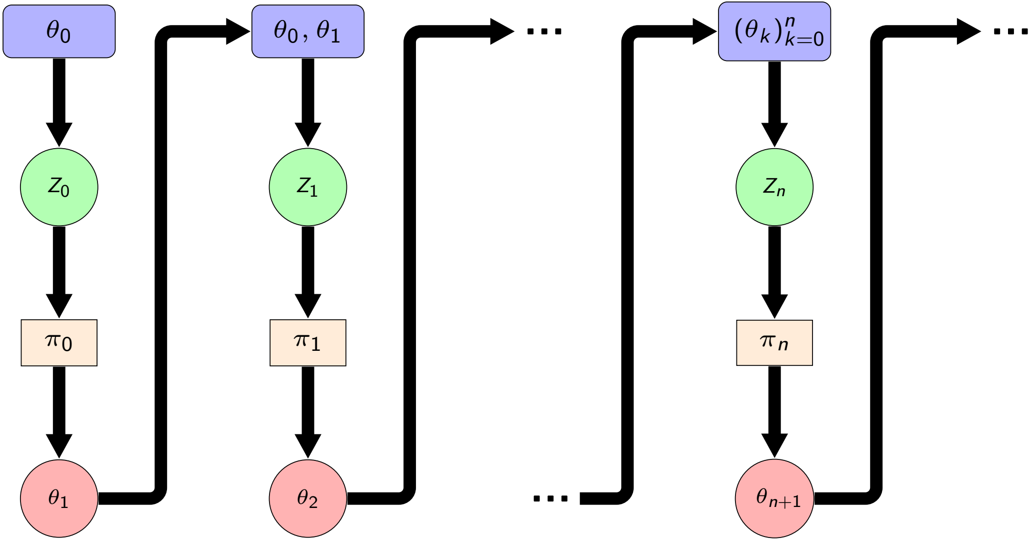
$$p_n(\theta|\bar{y}) = \frac{1}{C_{\bar{y}}^{(n)}} \frac{\varrho(\bar{y}|\theta)}{Z_n(\theta)},$$

with distribution $\pi_n(\mathrm{d}\theta) = p_n(\theta|\bar{y})\nu(\mathrm{d}\theta)$.

Having already computed $\theta_0, \theta_1 \ldots, \theta_n$ use the following scheme:

1) Compute (randomized) estimator Z_n for Z.
2) Sample (approximately) via MCMC from π_n; return result θ_{n+1}.

Illustration



Some properties of the process

Sequence $(\theta_n)_{n\in\mathbb{N}_0}$ provides a realization of a sequence of random variables $(\xi_n)_{n\in\mathbb{N}_0}$.

Some observations:

- Each ξ_{n+1} is related to kernel K_{Z_n} and limit distribution π_n.
- In each step kernel and limit distribution change.
- ξ_{n+1} may depend on the whole history $\xi_0, \ldots, \xi_n$.

Question: Can we use $(\xi_n)_{n\in\mathbb{N}_0}$ to approximate π (in some sense)?

Some known results

Given certain regularity conditions adaptive MCMC algorithms satisfy

- A strong law of large numbers for bounded functions Atchadé et al. (2013).
- A weak law of large numbers for integrable functions Liang et al. (2016).

That is, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} h(\xi_j) = \int_{\Theta} h(\theta)\pi(\mathrm{d}\theta),$$

almost surely, or in probability for a suitable class of functions.

Question: Can we have bounds of the error?

Theoretical results – some assumptions

We consider the following "regularity conditions":

- We assume that Z and the estimators Z_n are contained in a suitable class of functions $\mathscr{C}$.
- Given $\xi_n = y_n$ and $Z_n = z_n$ we assume that

$$\xi_{n+1} \sim K_{z_n}^{m_n}(y_n, \cdot).$$

- Assume that for any $z \in \mathscr{C}$ kernel K_z has invariant distribution π_z and that for any $n \in \mathbb{N}_0$ holds

$$\sup_{\theta \in \Theta} \left\| K_{Z_n}^{m_n}(\theta, \cdot) - \pi_n \right\|_{tv} \leq r,$$

  for some $r \in (0, 1)$.

Theoretical results – Error bound

Theorem

Assume the above regularity conditions and that for $n \in \mathbb{N}_0$ a.s. holds

$$\mathbb{E}|Z_n(\theta) - Z(\theta)|^2 \leq a_n \qquad \textit{and} \qquad \sup_{\theta \in \Theta} \left\| K_{Z_n}^{m_n}(\theta, \cdot) - \pi_n \right\|_{tv} \leq r_n,$$

where $(a_n)_{n\in\mathbb{N}_0}, (r_n)_{n\in\mathbb{N}_0} \in \mathbb{R}^{\mathbb{N}_0}$. Then, for any $n \in \mathbb{N}$ and $h \in L^\infty(\Theta, \nu)$ we have

$$\mathbb{E}\left| \frac{1}{n}\sum_{j=1}^{n} h(\xi_j) - \int_\Theta h(\theta)\pi(\mathrm{d}\theta) \right|^2 \leq C \left\|h\right\|_\infty^2 \left(\frac{1}{n} + \frac{1}{n}\sum_{j=1}^{n}(a_j + r_j^2) \right),$$

where $C \in (0, \infty)$ does not depend on h or n.

A nice consequence

The above result in particular yields a convergence rate which is uniform over all h with $\|h\|_\infty \leq 1$.

Corollary

Under the assumptions of the above theorem, there exists a constant $C \in (0, \infty)$ such that for any $n \in \mathbb{N}$ holds

$$\sup_{\|h\|_\infty \leq 1} \mathbb{E} \left| \frac{1}{n} \sum_{j=1}^{n} h(\xi_j) - \int_\Theta h(\theta)\pi(\mathrm{d}\theta) \right|^2 \leq C \left(\frac{1}{n} + \frac{1}{n} \sum_{j=1}^{n} (a_j + r_j^2) \right).$$

Theoretical results – CLT

Theorem

Assume the regularity conditions above and that for some $\alpha > 1/2$ *and* $n \in \mathbb{N}_0$ *holds*

$$\mathbb{E}|Z_n(\theta) - Z(\theta)|^2 \lesssim \frac{1}{n^\alpha} \qquad \text{and} \qquad \sup_{\theta \in \Theta} \left\| K_{Z_n}^{m_n}(\theta, \cdot) - \pi_n \right\|_{tv} \lesssim \frac{1}{n^\alpha}.$$

Then, for any measurable and bounded h*, with* $\sigma(h)^2 = \pi(h^2) - \pi(h)^2 \neq 0$ *we have*

$$\frac{1}{\sqrt{n}\sigma(h)} \sum_{j=1}^{n} (h(\xi_j) - \pi(h)) \longrightarrow \mathcal{N}(0, 1),$$

in distribution as $n \to \infty$.

Here $\pi(g) = \int_\Theta g(\theta)\pi(\mathrm{d}\theta)$.

Some remarks about our results

Our result shows that we can split the error of adaptive MCMC integration into:

- One part which matches the "iid rate".
- One part which is only depending on approximation error of Z.
- One part which reflects "how good" sampling from π_n is possible.

Note that the error is w.r.t. to the integral $\int_\Theta h(\theta)\pi(\mathrm{d}\theta)$, which in particular means we can have an asymptotic exact method.

The estimator(s) Z_n need not satisfy additional conditions such as unbiasedness or being iid (but the theorem also works if they do).

Some remarks about our results – computational costs

Computational costs of our method to get $\{\xi_1, \ldots, \xi_n\}$:

- The costs to obtain the Z_n's, which depend on the specific estimator one uses.
- The MCMC steps to sample from the π_n's, which are $\sum_{j=0}^{n-1} m_j$.
- In each MCMC step (for sampling from π_n) evaluation(s) of Z_n may be required which also could be "expensive".

The choice of $(m_n)_{n\in\mathbb{N}_0}$ as well as Z_n are problem specific, however, finding a "clever" estimator Z_n may pay off in the computational costs.

Main idea of the proofs

We split the Monte Carlo sum into three parts:

$$\sum_{j=1}^{n} h(\xi_j) - \pi(h) = M_n + R_1(n) + R_2(n),$$

with

- a martingale part M_n,
- R_1 only depending on $\mathbb{E}|Z_n(\theta) - Z(\theta)|^2$,
- R_2 only depending on $\sup_{\theta \in \Theta} \left\| K_{Z_n}^{m_n}(\theta, \cdot) - \pi_n \right\|_{tv}$.

Main idea of the proofs – II

Thereto we solve Poisson's equation: Given $h\colon \mathcal{Y} \to \mathbb{R}$ find u_γ such that

$$u_\gamma(y) - K_\gamma u_\gamma(y) = h(y) - \pi_\gamma(h).$$

Having computed the u_γ's we can use it within the sum

$$\sum_{j=1}^{n} \left(h(\xi_j) - \pi_j(h)\right),$$

which yields the desired martingale decomposition.

Example: Ising model

We consider an Ising model with $M_1 \times M_2$ nodes

$$\mathcal{X} = \left\{ \{x_{i,j}\}_{i=0,\ \ j=0}^{M_1-1,\, M_2-1} \colon x_{i,j} \in \{-1, 1\} \right\}$$

and energy function

$$E(x) = -\sum_{i=0}^{M_1-1} \sum_{j=0}^{M_2-1} x_{i,j} \cdot \left(x_{(i+1 \bmod M_1\,,j)} + x_{(i\,,j+1 \bmod M_2)}\right).$$

Set

- $\mathcal{Y} = \{y \colon \exists x \in \mathcal{X} \text{ with } y = E(x)\}$
- $\Theta = [0, K]$ for some $K < \infty$

Example: Ising model – II

We have $\varrho(y|\theta) = \exp(-y \cdot \theta)$ and

$$Z(\theta) = \sum_{x \in \mathcal{X}} \varrho(E(x)|\theta) \qquad \text{as well as} \qquad p(\theta|y) = \frac{1}{C_y} \frac{\varrho(y|\theta)}{Z(\theta)}.$$

For Z_n based on Importance Sampling and $h \in L^\infty(\Theta)$ we can show:

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{j=1}^{n} h(\xi_j) - \pi(h)\right|^2\right] \leq C \, \|h\|_\infty \, \frac{1}{n}$$

with $C \in (0, \infty)$ independent of n and also a CLT can be shown.

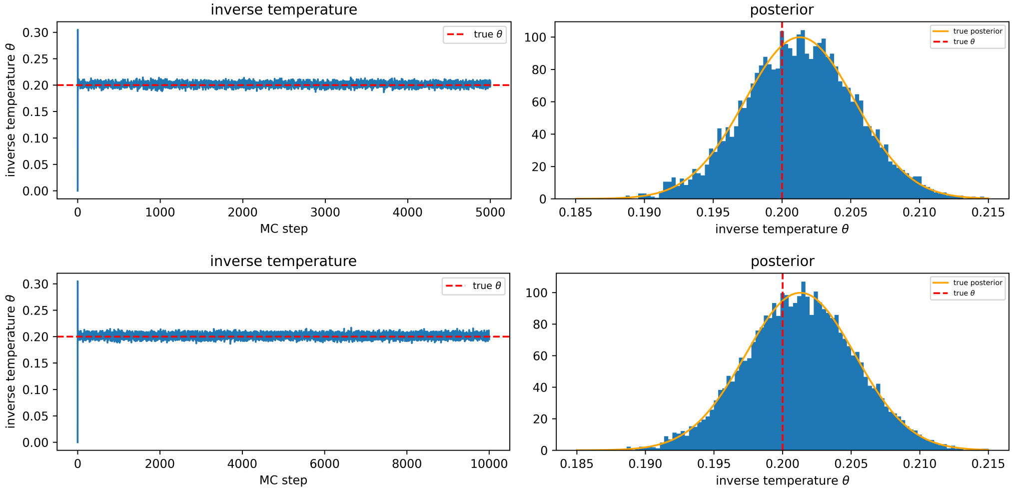
Example: Ising model – Setup for numerics

We did some simulations for the Ising model with the following parameters:

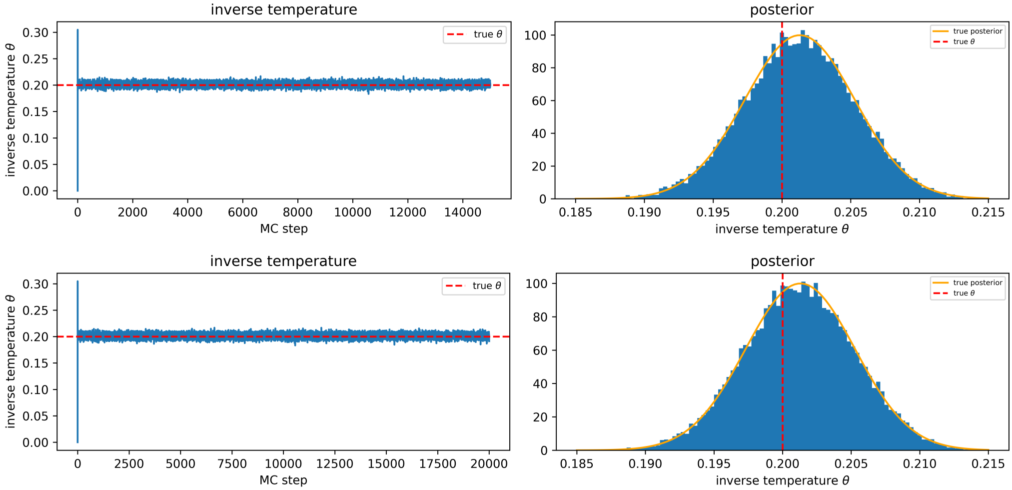
- Grid size: 16×16 (so 2^{256} summands in Z).
- Posterior is w.r.t. a measurement with true inverse temperature 0.2.
- Estimators Z_n are based on importance sampling.
- Choice of $m_n = 100 + (n \mod 250)$.

Time for computing a sample of size 30000: roughly 15 – 20 minutes.

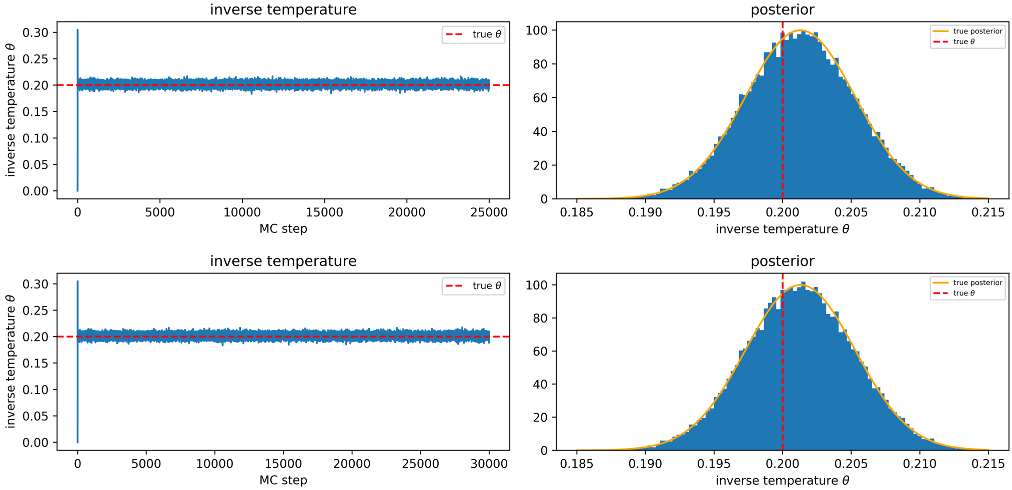
Example: Ising model – Numerics



Example: Ising model – Numerics II



Example: Ising model – Numerics III



Summary and outlook

Main message:

Explicit error bounds for adaptive MCMC within the doubly intractable setting are available.

Possible further topics to explore:

- Not only use the last step of the chains targeting π_n.
- Try to weaken the uniform ergodicity assumption.

Thank you for listening!

References I

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2016). Noisy monte carlo: Convergence of markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47.

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725.

Atchadé, Y. F., Lartillot, N., and Robert, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*, 27(4):416–436.

Eltzner, B., Hofstadler, J., Rudolf, D., Habeck, M., and de Groot, B. (2023). Bayesian maximum entropy ensemble refinement. *bioRxiv*.

Habeck, M. (2014). Bayesian approach to inverse statistical mechanics. *Physical Review E*, 89(5):052113.

References II

Habeck, M., Rudolf, D., and Sprungk, B. (2020). Stability of doubly-intractable distributions. *Electronic Communications in Probability*, 25:1–13.

Liang, F., Jin, I. H., Song, Q., and Liu, J. S. (2016). An adaptive exchange algorithm for sampling from distributions with intractable normalizing constants. *Journal of the American Statistical Association*, 111(513):377–393.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and computing*, 22(6):1167–1180.

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458.

References III

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). Mcmc for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 359–366. AUAI Press.

Park, J. and Haran, M. (2018). Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523):1372–1390.