# On cyclical MCMC sampling

**Liwei Wang**
Boston University

**Xinru Liu**
Boston University

**Aaron Smith**
University of Ottawa

**Yves Atchadé**
Boston University

## Abstract

Cyclical MCMC is a novel MCMC framework recently proposed by Zhang et al. (2019) to address the challenge posed by high-dimensional multimodal posterior distributions like those arising in deep learning. The algorithm works by generating a nonhomogeneous Markov chain that tracks – cyclically in time – tempered versions of the target distribution. We show in this work that cyclical MCMC converges to the desired probability distribution in settings where the Markov kernels used are fast mixing, and sufficiently long cycles are employed. However in the far more common settings of slow mixing kernels, the algorithm may fail to produce samples from the desired distribution. In particular, in a simple mixture example with unequal variance we show by simulation that cyclical MCMC fails to converge to the desired limit. Finally, we show that cyclical MCMC typically estimates well the local shape of the target distribution around each mode, even when we do not have convergence to the target.

## 1 INTRODUCTION

Over the last few decades, statistics and machine learning have become the dominant framework for scientific knowledge discovery and decision making from data. However uncertainty quantification remains one important aspect where further basic advancement is needed. In principle Bayesian statistics provides a coherent learning framework where uncertainty can be rigorously quantified. However, deploying the Bayesian machinery in practice invariably hinges on the ability to handle high-dimensional and often multimodal posterior distributions. Markov chain Monte Carlo (MCMC) is the state-of-the-art for dealing with this problem (Robert and Casella (2004); Douc et al. (2018)). Despite several decades of progress in MCMC, sampling from multimodal distributions, particular in a high-dimensional context, remains extremely challenging. The state-of-the-art for dealing with multimodality is the idea of "tempering" - that is, building a sequence of distributions that bridges the target distribution and some other distribution that is easier to sample. Simulated tempering (ST) and parallel tempering (PT) are the two main algorithms built on this principle (Geyer (1991); Marinari and Parisi (1992); Hukushima and Nemoto (1996); Geyer and Thompson (1995)). However to maintain correctness, these algorithms require additional auxiliary variables and Metropolis steps that significantly increase their implementation costs.

In Zhang et al. (2019) the authors proposed "cyclical stochastic gradient MCMC", a fast implementation of tempering, that dispenses with the costly auxiliary variables required in ST and PT. The algorithm implements tempering without any Metropolis step in the temperature dimension. The algorithm operates in cycles, and is designed such that in the initial part of a cycle there is an exploration of the space to find a mode, and in the second part of a cycle, samples are drawn from the mode found. In this paper we use the generic term "cyclical MCMC" to refer to algorithms built on that principle. The purpose of this work is to analyze these algorithms.

We found that in general cyclical MCMC does not converge to the correct target distribution. For instance, we show by simulation that in a simple Gaussian mixture model with unequal variance, cyclical MCMC does not correctly recover the weights of the mixture.

Using a novel adaptation of the spectral gap technique to nonhomogeneous Markov chains, we show that cyclical MCMC does converge to the intended target distribution when the Markov kernels used have fast enough mixing, and an appropriate tempering schedule is selected. We also found that even when the weights of the mixture are poorly approximated,

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

cyclical MCMC typically produces a correct approximation of the shape of the distribution around each mode.

The remaining of the paper is organized as follows. We provide some motivating background in Section 2. The cyclical MCMC algorithm is described in Section 3. Our theoretical results are described in Section 4, with most proofs collected in Section 6 and in the supplement.

## 2 MOTIVATION: BAYESIAN INFERENCE

Suppose that we have data $\mathcal{D} \stackrel{\text{def}}{=} \{(\mathbf{x}_i, \mathbf{y}_i),\ 1 \le i \le n\}$, where $\mathbf{y}_i|\mathbf{x}_i \sim f_\theta(\mathbf{y}_i|\mathbf{x}_i)$, for some statistical model $f_\theta$ with (vectorized) parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Assuming independence, the log-likelihood writes

$$\ell(\theta; \mathcal{D}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell_i(\theta), \quad \text{where} \quad \ell_i(\theta) \stackrel{\text{def}}{=} \log f_\theta(\mathbf{y}_i|\mathbf{x}_i).$$

In the Bayesian paradigm we complement the model with a prior distribution $\mu_0$ on $\theta$ that summarizes all available prior knowledge (our inductive bias), such as boundedness, smoothness, sparsity, etc. The resulting posterior distribution of $\theta$ is the probability measure

$$\Pi(\mathrm{d}\theta|\mathcal{D}) \propto e^{\ell(\theta;\mathcal{D})}\mu_0(\mathrm{d}\theta). \tag{1}$$

$\Pi(\cdot|\mathcal{D})$ captures the uncertainty in the estimation of $\theta$. Given a new data point for which we observe $\mathbf{x}$, we can predict corresponding response $\mathbf{y}$ by sampling from the posterior predictive distribution

$$f(\cdot|\mathbf{x}) \stackrel{\text{def}}{=} \int_{\mathbb{R}^p} f_\theta(\cdot|\mathbf{x})\Pi(\mathrm{d}\theta|\mathcal{D}). \tag{2}$$

The uncertainty in a draw from $f$ is the combination of the uncertainty in learning the model, as reflected by the posterior distribution, plus the inherent uncertain of the model itself. Some authors (Hüllermeier and Waegeman (2021)) use the terms epistemic and aleatoric uncertainty respectively to refer to these two sources of uncertainties.

In practice, the posterior integral in (2) is rarely tractable, and so it is often replaced by an approximation

$$\hat{f}(\cdot|\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{K}\sum_{k=1}^{K} f_{\theta^{(k)}}(\cdot|\mathbf{x}), \tag{3}$$

where $\{\theta^{(k)},\ 1 \le k \le K\}$ are approximately sampled from $\Pi(\cdot|\mathcal{D})$. Taking $K = 1$, and $\theta^{(1)}$ as the estimate obtained by running stochastic gradient descent (SGD) provides a very poor representation of the epistemic uncertainty. Other approaches such as variational approximation (Graves (2011); Blei et al. (2016)) or dropout (Srivastava et al. (2014); Gal and Ghahramani (2016)) produce better approximations of $\Pi(\cdot|\mathcal{D})$, but still, are known to misrepresent the epistemic uncertainty. At the other end of the spectrum, traditional MCMC methods produce (asymptotically) correct samples, but are computationally too expensive in many large applications. There is therefore a pressing need for methods, such as cyclical MCMC, that aim to strike a better balance between cost and accuracy. However a good theoretical understanding of these algorithms is needed.

## 3 CYCLICAL MCMC SAMPLING

Let $\Theta \subseteq \mathbb{R}^p$ be the state space. Suppose that we are interested in a probability density $\Pi$ (with respect to the Lebesgue measure) of the form

$$\Pi(\theta) = \frac{\exp\left(-\mathcal{E}(\theta)\right)}{Z}, \quad \theta \in \Theta, \tag{4}$$

where $Z$ is the normalizing constant, and $\mathcal{E} : \Theta \to \mathbb{R}$ some arbitrary measurable function. Let $\beta : [0,1] \to [0,1]$ be a continuously differentiable function with $\beta(0) = \beta(1) = 1$. We also assume that $\beta$ is decreasing on $[0, 1/2]$, and increasing on $[1/2, 1]$. We extend $\beta$ into a function $\beta :\ [0,\infty) \to [0,1]$ by period extension (meaning that for all $x \ge 0$, $\beta(x+1) = \beta(x)$). For example, following (Zhang et al. (2019)) we consider in our simulations the choice[1]

$$\beta(t) = \frac{1 + \cos(2\pi t^r)}{2}, \quad t \ge 0 \tag{5}$$

for some power $r \ge 1$.

Let $L \ge 1$ be a cycle length. For integer $j \ge 1$, and with $\beta_j \stackrel{\text{def}}{=} \beta(j/L)$, we define the density

$$\Pi_j(\theta) = \frac{1}{Z_j}\exp\left(-\beta_j\mathcal{E}(\theta)\right), \quad \theta \in \Theta. \tag{6}$$

Since $\beta$ is periodic with period 1 and $\beta(0) = 1$, it holds that $\Pi_L = \Pi_{2L} = \cdots \Pi_{kL} = \Pi$, for all $k \ge 1$. Furthermore, as $j$ increases from 1 to $L/2$, the distribution $\Pi_j(\mathrm{d}\theta)$ becomes more diffuse, and its shape is restored back to $\Pi$ as $j$ increases from $L/2$ to $L$.

For each $j \ge 1$, let $M_j$ be a Markov kernel on $\Theta$ that can be used to sample from $\Pi_j$. In the theoretical investigation we will assume that $M_j$ has invariant distribution $\Pi_j$. In practice, Markov kernels that do not maintain $\Pi_j$ as invariant distribution (such as stochastic gradient Langevin dynamics (SGLD)) are used, but

[1] In the simulations we actually use $\max(\beta(t), 0.001)$ to prevent $\beta(t) = 0$ which may be problematic when $\Theta$ is unbounded.

we do not analyze these here. Given some initial distribution $\nu^{(0)}$, the idea of cyclical MCMC as developed in (Zhang et al. (2019)) consists in simulating the nonhomogeneous Markov chain $\{\theta^{(j)},\ j \geq 0\}$, where $\theta^{(0)} \sim \nu^{(0)}$, and for $j \geq 1$,

$$\theta^{(j)} \mid \{\theta^{(0)}, \ldots, \theta^{(j-1)}\} \sim M_j(\theta^{(j-1)}, \cdot). \qquad (7)$$

The chain is run for $K$ cycles (meaning for $K \times L$ iterations), and we collect the samples obtained at (or around) the end of each cycle to form $\{\theta^{(kL)},\ 1 \leq k \leq K\}$ as our sample representation of $\Pi$. A synoptic view of the algorithm is given in Algorithm 1.

**Remark 1.** The intuition of the method when $\Pi$ is multimodal is that, as the kernel $M_j$ changes along the cycle and targets $\Pi_j$, by the middle of the cycle (since $\Pi_{1/2}$ is more diffuse), the sampler is able to escape more easily from any current (local) mode. However by the end of the cycle as $\beta_j$ moves closer to 1, the algorithm returns back to targeting $\Pi$. Hence, there is an exploration phase with mode discovery, followed by an exploitation phase where samples are drawn from around the selected mode. □

**Remark 2.** The kernel $M_j$ can be constructed using any standard MCMC algorithm that is applicable (Gibbs sampling, Metropolis-Hastings, HMC, etc...). We note however, since the distributions $\Pi_j$ have markedly different covariance structures, for good performance it is important to properly scale the proposal kernels accordingly. In Zhang et al. (2019) the authors mostly focused on SGLD (Welling and Teh (2011); Raginsky et al. (2017)) and SGHMC (Chen et al. (2014); Ma et al. (2015)). □

**Algorithm 1.** [Cyclical MCMC]
Choose the function $\beta : [0,1] \to [0,1]$, the cycle length $L$, the number of cycle $K$, the initial distribution $\nu^{(0)}$, and construct the sequence of nonhomogeneous Markov kernel $\{M_j,\ j \geq 1\}$.

1. Draw $\theta^{(0)} \sim \nu^{(0)}$.
2. For $j = 1, \ldots, K \times L$, draw
$$\theta^{(j)} \mid \{\theta^{(0)}, \ldots, \theta^{(j-1)}\} \sim M_j(\theta^{(j-1)}, \cdot).$$
3. Return $\{\theta^{(kL)},\ 1 \leq k \leq K\}$.

### 3.1 A toy example

Although the idea of cyclical MCMC is intuitively clear and appealing, the algorithm typically does not converge to the correct limit. For instance, consider a simple one-dimensional mixture density

$$\Pi(\theta) = \frac{1}{2} f_1(\theta) + \frac{1}{2} f_2(\theta).$$

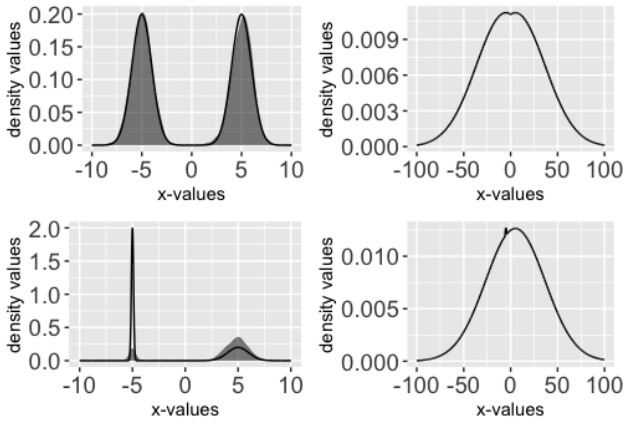


Figure 1: Top left (resp. bottom left) is the density $\Pi$ and its estimate as produced by cyclical MCMC when $c = 1$ (resp $c = 0.1$). Top right (resp. bottom right) shows the powered density $\Pi^{0.001}$ with $c = 1$ (resp $c = 0.1$).

where $f_1$ (resp. $f_2$) is the density of $\mathbf{N}(5,1)$ (resp. $\mathbf{N}(-5, c^2)$) where $c$ is either $c = 1$ (equal variance mixture) or $c = 0.1$ (unequal variance mixture). We apply Algorithm 1 to sample from $\Pi$ where $M_j$ is taken as a random walk Metropolis with proposal $\mathbf{N}(x, 0.25/\beta_j)$. The initial distribution is $\mathbf{N}(0,1)$. We use $K = 1000$, $L = 5000$, and $\beta$ as in (5) with $r = 1$. In Figure 1, the shaded areas on the left side depict the densities estimated from the cyclical MCMC outputs, whereas the solid black lines on the left side represent the true density. In the top-left plot, where $c = 1$ (equal variance), the recovery is excellent and it is hard to distinguish the two curves. In the bottom-left where $c = 0.1$ (unequal variance), the recovery is poor. In the unequal variance setting cyclical MCMC systematically fails to correctly estimate the weights of the mixture. In this specific example, the estimate of the weight of $f_1$ is 0.497 in the equal variance setting – which is excellent, but 0.87 in the unequal variance setting.

The issue in this example is that in the unequal variance setting, tempering by powering leads to a witch-hat distribution (Matthews (1993)), with a small bulge at the top (bottom-right side of Figure 1) that is virtually impossible to detect by most MCMC algorithms – but is crucial to properly approximate the weights of the mixture. As a result the algorithm spends more time on the positive side of the line and produces an estimate of the mixture weights that is biased towards $f_1$. In the equal variance setting, powering produces a nice smooth symmetric density (top-right side of Figure 1) and the corresponding Markov kernels $M_j$'s are fast mixing. Parallel tempering and simulated tempering are known to suffer from the same poor mixing in

the unequal variance setting (Woodard et al. (2009a)). However unlike cyclical MCMC, these algorithms have a built-in Metropolis-Hastings acceptance step that restores the correctness of the target distribution.

# 4 SOME THEORETICAL INSIGHTS

In this section we analyze the convergence of cyclical MCMC. First, we analyze the behavior of Algorithm 1 in the case where the Markov kernels $M_j$'s are fast mixing. In that setting we show that indeed cyclical MCMC can converge to $\Pi$. In the second part we consider a more realistic setting where the cycle of the algorithm can be decomposed into a mode exploration part, and a mode exploitation part. In that regime we show that cyclical MCMC converges to a mixture with correct mixture components, but with possibly incorrect weights.

## 4.1 The fast mixing regime

Since the algorithm generates a nonhomogeneous Markov chain, classical Markov chain mixing time analysis do not apply. We extend the classical Markov chain spectral gap theory to handle nonhomogeneous Markov chains.

Let $\mathcal{B}_\Theta$ denote the Borel sigma-algebra of $\Theta$. We recall that the target density $\Pi$ and the sequence of densities $\Pi_j$ are defined in (4) and (6) respectively. We will abuse notation and also write $\Pi$ (resp. $\Pi_j$) to denote the probability measure on $\Theta$ with density $\Pi$ (resp. $\Pi_j$). For instance we will write $\Pi(A)$ as a short for $\int_A \Pi(\theta)\mathrm{d}\theta$.

To proceed, we need some Markov chain notations. A good reference is (Meyn and Tweedie (2009)). Given two Markov kernels $Q_1, Q_2$ on $\Theta$, their product is the Markov kernel $Q_1Q_2$ defined[2] as $(Q_1Q_2)(u,A) \stackrel{\text{def}}{=} \int_\Theta Q_1(u,\mathrm{d}v)Q_2(v,A)$. This multiplication can naturally be iterated. Given a Markov kernel $Q$, and a probability $\nu$, the product $\nu Q$ denotes the probability measure $(\nu Q)(A) \stackrel{\text{def}}{=} \int_\Theta \nu(\mathrm{d}u)Q(u,A)$. Furthermore, given a measure $\nu$ on $\Theta$, and a function $h:\ \Theta \to \mathbb{R}$, we write $\nu(h) \stackrel{\text{def}}{=} \int_\Theta h(u)\nu(\mathrm{d}u)$. For $j \geq 1$, we let

$$\mathcal{L}_j^2 \stackrel{\text{def}}{=} \mathcal{L}^2(\Pi_j) \stackrel{\text{def}}{=} \left\{h:\ \Theta \to \mathbb{R}:\ \Pi_j(h^2) < \infty\right\},$$

and for $h, h_1, h_2 \in \mathcal{L}_j^2$, we set $\mathsf{Var}_j(h) \stackrel{\text{def}}{=} \Pi_j(h^2) - \Pi_j(h)^2$ and $\langle h_1, h_2\rangle_j \stackrel{\text{def}}{=} \int_\Theta h_1 h_2 \mathrm{d}\Pi_j$. Given two finite measures $\mu, \nu$ on $\Theta$ the total variation distance between $\mu$ and $\nu$ is

$$\|\mu - \nu\|_{\text{tv}} \stackrel{\text{def}}{=} \sup_{f:\,|f|\leq 1} |\mu(f) - \nu(f)|.$$

Given a function $h:\ \Theta \to \mathbb{R}$, we set $\|h\|_\infty = \sup_{u\in\Theta}|h(u)|$. We impose the following assumption.

**H 1.** *For all $1 \leq j \leq L$, the Markov kernel $M_j$ is reversible with respect to $\Pi_j$, and*

$$\|\Pi_{j-1}/\Pi_j\|_\infty < \infty. \tag{8}$$

**Remark 3.** Reversibility is imposed here for convenience, and can be removed by introducing the adjoints of the $M_j$'s. Reversibility is a commonly imposed assumption in Markov chain theory, and is satisfied by many MCMC samplers. For instance all Metropolis-Hastings samplers, by definition generate reversible Markov kernels. However there are many other MCMC algorithms that are not reversible. Extending our results beyond the reversible case is an important question for future research.

The boundedness assumption $\|\Pi_{j-1}/\Pi_j\|_\infty < \infty$ is typically satisfied when $\Theta$ is bounded, and can essentially be viewed as assuming that $\Theta$ is bounded. Extending our framework to remove that assumption is an important direction for future work. □

The marginal distribution of $\theta^{(j)}$ at the $j$-th iteration of cyclical MCMC is given by

$$\nu^{(j)} \stackrel{\text{def}}{=} \nu^{(0)}M_1 \times \cdots \times M_j.$$

We seek conditions under which

$$\nu^{(kL)} \approx \Pi.$$

In what follows we set

$$\alpha_j \stackrel{\text{def}}{=} \sup\left\{\Delta_j(f),\ \ f \in \mathcal{L}_{j-1}^2,\ \Pi_{j-1}(f^2) = 1\right\},$$

where

$$\Delta_j(f) \stackrel{\text{def}}{=} \int_\Theta \Pi_{j-1}(\mathrm{d}x)f^2(x)\int_\Theta M_j^2(x,\mathrm{d}z)\left|\frac{\Pi_{j-1}(z)}{\Pi_j(z)} - 1\right|.$$

$\alpha_j$ is a measure of discrepancy between $\Pi_{j-1}$ and $\Pi_j$ and can be controlled by the choice of the function $\beta$ and the cycle length $L$. The next result illustrates this point. The proof is given in the supplement.

**Proposition 4.** *Suppose that $\Theta$ is bounded, and*

$$\mathsf{osc}(\mathcal{E}) \stackrel{\text{def}}{=} \max_{x,y}|\mathcal{E}(x) - \mathcal{E}(y)|,$$

*is finite, where $\mathcal{E}$ is as in (4). Then there exists a constant $C$, such that for all $1 \leq j \leq L$*

$$\alpha_j \leq \frac{C|\dot\beta(t_j)|}{L}, \tag{9}$$

*for some $t_j$, with $(j-1)/L \leq t_j \leq j/L$, where $\dot\beta$ is the derivative of $\beta$.*

[2] In this definition we follow the convenient practice in Markov chains theory of writing the integrant after the integrating measure.

*Proof.* See Section 9 of the Supplement. □

**Remark 5.** A more natural measure of discrepancy between $\Pi_{j-1}$ and $\Pi_j$ is their total variation distance

$$\|\Pi_{j-1} - \Pi_j\|_{\text{tv}} = \int_\Theta \left| \frac{\Pi_{j-1}(z)}{\Pi_j(z)} - 1 \right| \Pi_j(z)\mathrm{d}z.$$

However the total variation distance does not appear naturally in our analysis. □

For each $j \geq 1$, let $Q_j : \Theta \times \mathcal{B}_\Theta \to \mathbb{R}$ be the finite kernel defined as

$$Q_j(x, A) \stackrel{\text{def}}{=} \int_\Theta M_j(x, \mathrm{d}y) \int_A M_j(y, \mathrm{d}z) \frac{\Pi_{j-1}(z)}{\Pi_j(z)}. \tag{10}$$

Note that for each $x \in \Theta$, $Q_j(x, \cdot)$ is a finite measure on $(\Theta, \mathcal{B}_\Theta)$, but not necessarily a probability measure. We show in Section 9 of the supplement that under H1, $Q_j$ induces an operator $Q_j : \mathcal{L}^2_{j-1} \to \mathcal{L}^2_{j-1}$ that is self-adjoint and positive. We define the "spectral gap" of the operator $Q_j$ as

$$\lambda_j \stackrel{\text{def}}{=} \inf \left\{ \mathsf{G}_j(f),\ f \in \mathcal{L}^2_{j-1},\ \mathsf{Var}_{j-1}(f) > 0 \right\},$$

where

$$\mathsf{G}_j(f) \stackrel{\text{def}}{=} \frac{\int_\Theta \int_\Theta \left(f(y) - f(x)\right)^2 \Pi_{j-1}(\mathrm{d}x) Q_j(x, \mathrm{d}y)}{\int_\Theta \int_\Theta \left(f(y) - f(x)\right)^2 \Pi_{j-1}(\mathrm{d}x) \Pi_{j-1}(\mathrm{d}y)}.$$

**Remark 6.** $\lambda_j$ is large when the spectral gap of $M_j$ is large. This is easily seen from the definition of $Q_j$. One can also easily show for example that if for all $x$, $\|M_j(x, \cdot) - \Pi_j\|_{\text{tv}} \leq 2(1 - \beta_j)$ for some $\beta_j > 0$ that can be viewed as the spectral gap of $M_j$, then $\lambda_j \geq \beta_j$. This is because the last total variation norm inequality is equivalent to $M_j(x, \cdot) \geq \beta_j \Pi_j(\cdot)$ for all $x$, which in turn implies that $Q_j(x, \cdot) \geq \beta_j \Pi_{j-1}(\cdot)$, and so $\lambda_j \geq \beta_j$.

Finite kernels that are not necessarily Markov appear commonly in Markov chain theory (for instance in the analysis of sequence Monte Carlo samplers or in large deviations for Markov chains; see e.g. Kontoyiannis et al. (2006); Hervé (2008); Whiteley (2013)). □

With $\alpha_0 \stackrel{\text{def}}{=} 0$, we set

$$\Lambda_L \stackrel{\text{def}}{=} \sum_{i=0}^{L} \alpha_i \prod_{\ell=i+1}^{L} (1 - \lambda_\ell + \alpha_\ell).$$

Under assumption H1 we show in (23) that $\lambda_j \leq 1 + \alpha_j$. Hence it is always true that $\Lambda_L \geq 0$. Our main result of this section is as follows.

**Theorem 7.** *Assume H1. Let* $\nu^{(0)}(\mathrm{d}x) = f_0(x)\Pi(\mathrm{d}x)$. *Then for all* $k \geq 1$,

$$\|\nu^{(kL)} - \Pi\|^2_{\text{tv}} \leq \mathsf{Var}_0(f_0) \Lambda_L^k.$$

*Proof.* See Section 6. □

**Remark 8.** To explore the implications of this result, suppose for instance that

$$\lambda_j \geq \underline{\lambda} > 0, \quad \text{for all } \ j \geq 1. \tag{11}$$

Then using Proposition 4, we see that we can choose the tempering $\beta$ and $L$ such that $\alpha_j \leq C|\dot{\beta}(t_j)|L^{-1} \leq \underline{\lambda}/2$. In that case we have

$$|\Lambda_L| \leq \frac{C}{L} \sum_{i=0}^{L} |\dot{\beta}(t_i)| \left(1 - \frac{\underline{\lambda}}{2}\right)^{L-i} \leq \frac{2C \times \|\dot{\beta}\|_\infty}{L\underline{\lambda}} < 1,$$

for $L > 2C \times \|\dot{\beta}\|_\infty / \underline{\lambda}$. The condition (11) is a fast mixing condition on the kernels $\{M_j,\ j \geq 1\}$. We note that a similar but more subtle analysis can also be developed in cases where some of the initial Markov kernels $M_j$ (for $j$ closed to 1) have poor mixing. □

**Remark 9.** Our theorem and the discussion above thus show that in settings where all the Markov kernels $M_j$ have fast mixing, cyclical MCMC does converge to the right target distribution. For instance, in the mixture density example with equal variance setting, it is well-known that tempering by powering significantly improve mixing (Woodard et al. (2009b)). Thus we expect cyclical MCMC to work well for $L$ large enough, and this is what we observed in the simulations. In contrast, in the unequal variance setting, it is also known that powering does not improve mixing (Woodard et al. (2009a)). Hence in that case all the $\lambda_j$'s remain close to 0, and cyclical MCMC would require exponentially large cycle length $L$ to work. □

**Remark 10.** One natural objection to Theorem 7 is this: for many MCMC problems where the tails of $\Pi$ are poorly understood and $\nu^{(0)}$ badly chosen, the variance term $\mathsf{Var}_0(f_0)$ appearing in the conclusion of Theorem 7 is infinite, and so the theorem does not give nontrivial upper bounds. Fortunately, in many examples, this problem can be easily fixed via truncation as we explain in Section 10 of the supplement. □

### 4.2 The highly multimodal regime

We show here that even when it fails to capture correctly the weights of the mixture, cyclical MCMC typically estimates well the component densities in the mixture. In this section we assume that $\Pi$ is a mixture of the form

$$\Pi(\theta) = \sum_{i=1}^{d} w_i f_i(\theta). \tag{12}$$

Let $\Theta_1, \ldots, \Theta_d \subset \Theta$ be a collection of disjoint subsets of $\Theta$ such that $\Theta_j$ contains the bulk of the probability

mass of $f_j$. For each $j \in [d]$, let $I_j \subset \Theta_j$, where $[d]$ is a short for $\{1, \ldots, d\}$. Finally, fix $L_2 \in [L]$. We consider the process $\{\theta^{(j)},\ j \geq 0\}$ generated by Algorithm 1 applied to (12).

The main result in this section, Theorem 13, has three main error terms. Immediately after each term is introduced, we verify that it is small for Metropolis-Hastings chains with ultimate target

$$\Pi = \frac{1}{2}\mathbf{N}(-1, \sigma^2) + \frac{1}{2}\mathbf{N}(-1, c^2\sigma^2) \tag{13}$$

for $0.5 < c < 2$, and proposal kernel $Unif([\theta - \frac{i}{L}\sigma, \theta + \frac{i}{L}\sigma])$. These simple chains are similar to the toy example considered in Section 3.1.

Our first error term comes from the following "no-escape" assumption:

**H 2.** *There exists some $0 \leq \delta_1 < 1$ with the following property. For any $j \in [d]$ and any $\theta \in I_j$, the nonhomogeneous Markov chain*

$$\theta^{(L_2)} = \theta,\ \ \theta^{(i)} \sim M_i(\theta^{(i-1)}, \cdot) \qquad L_2 < i \leq L \tag{14}$$

*satisfies*

$$\mathbb{P}\left(\bigcup_{i=L_2+1}^{L} \{\theta^{(i)} \notin \Theta_j\}\right) \leq \delta_1. \tag{15}$$

**Remark 11.** One can verify Assumption 2 by showing that the kernels $M_j$ satisfy a Lyapunov drift condition. We explain the details in Section 12 of the supplement. For the target given in Equation (13), for $\sigma$ small enough, we take $\Theta_1 = [-1.5, -0.5]$ and $\Theta_2 = [0.5, 1.5]$ and time $\frac{L}{3} \leq L_2 \leq \frac{2L}{3}$, and the existence of a Lyapunov drift condition is given by Lemma 17 of the supplement. □

For any $j \in [d]$, denote by $M_i^{(j)}$ the Metropolis-Hastings chain with proposal distribution $M_i$ and target distribution $\Pi_i^{(j)}$ with density proportional to $\Pi_i(\theta)\mathbf{1}_{\theta \in \Theta_j}$; we call this the "restriction" of $M_i$ (respectively $\Pi_i$) to the set $\Theta_j$. We make the following "mixing within modes" assumption:

**H 3.** *There exists $0 \leq \delta_2 < 1$ with the following property. Fix $j \in [d]$ and $\theta \in I_j$. Define the time-inhomogeneous Markov chain:*

$$\theta^{(L_2)} = \theta,\ \theta^{(i)} \sim M_i^{(j)}(\theta^{(i-1)}, \cdot) \qquad L_2 < i \leq L. \tag{16}$$

*This chain satisfies:*

$$\|\mathbb{P}(\theta^{(L)} \in \cdot) - \Pi^{(j)}(\cdot)\|_{\mathsf{tv}} \leq \delta_2. \tag{17}$$

**Remark 12.** The inequality in Assumption 3 is exactly the conclusion of Theorem 7 but applied to the restricted process, and so that theorem can be used to verify Assumption 3. □

Under these assumptions, we have:

**Theorem 13.** *Let Assumptions 2, 3 hold and let $\delta = \delta_1 + \delta_2$. Then the Markov chain satisfies:*

$$\|\nu^{(L)} - \sum_{j=1}^{d} \nu^{(L_2)}(I_j)\Pi^{(j)}\|_{\mathsf{tv}} \leq \delta + (1 - \nu^{(L_2)}(\cup_{j=1}^{d} I_j)). \tag{18}$$

*Proof.* See Section 7. □

**Remark 14.** The remainder term $(1 - \sum_{j=1}^{d} \nu^{(L_2)}(I_j))$ appearing in Theorem 13 can be bounded using the same Lyapunov function approach described in Remark 11. In particular, assume there exists a function $V : \Theta \to [1, \infty)$ and constants $m > 0$, $1 < L_1 < L_2$, $0 < \alpha \leq 1$, and $0 \leq \beta < \infty$ with following properties:

1. For $\theta \notin \cup_j \Theta_j$, we have $V(\theta) \geq e^m$.
2. For all $\theta \in \cup_j \Theta_j$ and all $L_1 \leq i \leq L_2$, one has
$$M_i V(\theta) \leq (1 - \alpha)V(\theta) + \beta.$$

Then applying Markov's inequality, we find

$$\mathbb{P}\left(X_{L_2} \notin \cup_j \Theta_j\right) \leq \mathbb{P}\left(V(X_{L_2}) > e^m\right) \leq e^{-m}\mathbb{E}\left(V(X_{L_3})\right) \leq e^{-m}\left(\frac{\beta}{\alpha} + \mathbb{E}\left(V(X_{L_1})\right)\right).$$

Such a Lyapunov function is known to exist in substantial generality. For example, see Theorem 4.1 of Jarner and Hansen (2000) for conditions under which $V(\theta) = \Pi(\theta)^{-c}$ is a Lyapunov functions for $M_L$ for all $0 < c < 1$; this result covers the example in Equation (13). If we choose $L_1 \geq aL$, this means that for all $0 < c < \frac{a}{2}$ the single function $V(\theta) = \Pi(\theta)^{-c}$ is simultaneously a Lyapunov function for all chains $M_i$ with $L_1 \leq i \leq L_2$. □

**Remark 15.** A reviewer insightfully asked whether Theorem 13 can be extending to analyze $\nu^{(kL)}$. This can be done using the following general argument. If a Markov kernel $P$ and a probability measure $\nu$ (not nec. invariant under $P$) satisfy $\|P(x, \cdot) - \nu\|_{\mathsf{tv}} \leq \rho$, for all $x$, and for some $\rho \in (0, 1)$, then we can say that $P$ satisfies Doeblin's condition, and therefore admits an invariant $\pi$, say, and it holds $\|\pi - \nu\|_{\mathsf{tv}} \leq \rho$. Therefore, for all $k \geq 1$, $\|P^k(x, \cdot) - \nu\|_{\mathsf{tv}} \leq \kappa^k + \rho$, for some $\kappa \in (0, 1)$. □

#### 4.2.1 Illustration with a two-dimensional mixture

As illustration of Theorem 13 we consider a two-dimensional Gaussian mixture with 25 components

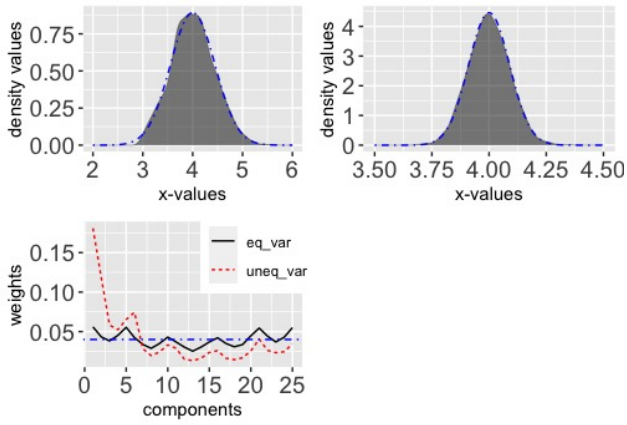


Figure 2: Top row is the $x$-component of the component density $f_{25}$ and its estimate as produced by cyclical MCMC. The true densities are plotted in blue dashed line. Top-left is the equal variance setting. Bottom left shows the estimated weights of the mixture.

adapted from Zhang et al. (2019), where

$$\Pi(\theta) = \frac{1}{25}\sum_{i=1}^{25} f_i(\theta),\ \theta \in \mathbb{R}^2,$$

where $f_i$ is the density of the Gaussian distribution $\mathbf{N}(\mu_i, \sigma_i^2 I_2)$. The means $\mu_i$'s are the elements of $\{-4,-2,0,2,4\} \times \{-4,-2,0,2,4\}$, and $\sigma_i^2 = 0.2$ in the equal variance setting, and $\sigma_i^2 = 0.2/i$, $1 \le i \le 25$ in the unequal variance setting. In both cases we apply Algorithm 1 with $L = 20,000$, $K = 50,000$ using a random walk Metropolis with proposal $\mathbf{N}(x, 0.01\beta_j^{-1/2} I_2)$, and we take $\beta$ as in (5) with $r = 1$ (we obtained similar results for values of $r \le 20$ that we tried). The results are shown on Figure 2. The estimate of the mixture weights (a uniform distribution on $\{1,\dots,25\}$) is given on the second row of Figure 2. Again, we see that in the constant variance setting we recover correctly the weights, but the recovery is poor in the unequal variance setting.

We also look at the $x$-component of $f_{25}$. The shaded area in the top-left (resp. top-right) plot of Figure 2 shows the estimated density in the equal (resp. the unequal variance) variance setting. The true densities are plotted in dashed-line. We see that in both settings the recovery is good, which confirms the conclusion of Theorem 13.

## 5 CONCLUSION REMARKS

Cyclical MCMC is an efficient implementation of tempering that operates without the costly auxiliary variables required in PT and ST. The relationship between Cyclical MCMC and ST/PT is similar in some sense to the relationship between the Unadjusted Langevin algorithm and the Metropolis-Adjusted Langevin algorithm Durmus et al. (2019).

An initial analysis of the Cyclical MCMC was undertaken by Zhang et al. (2019). However their convergence upper bounds are actually non-informative. We show in this work that the algorithm does converge to the desired limit in settings where the Markov kernels used are fast mixing, and sufficiently long cycles are used. However in the far more common settings of slow mixing kernels, the algorithm may fail to converge to the correct limit. Indeed, in a simple mixture example with unequal variance where tempering is known to produce slow mixing kernels (Woodard et al. (2009a)), we show by simulation that cyclical MCMC fails to converge to the desired limit. However on the bright side, we also found that even when it fails to capture correctly the weights of the modes, cyclical MCMC typically estimates well the local shape of each mode.

Since it is biased toward flatter modes, in the context of Bayesian inference (Section 2), it appears that cyclical MCMC leads to a systematic over-estimation of the epistemic uncertainty, which may be a desirable feature in a prediction setting. It has also been observed empirically that in deep learning, flatter modes seem generalize better (see e.g. Pittorino et al. (2021)). It is thus possible that the biased asymptotic behavior of cyclical MCMC that we identified here may become a useful feature in some settings, although more research is needed on this issue.

The bias of the algorithm that we identified in the unequal variance mixture examples is ultimately a limitation of powering as a way of tempering. Our work thus also raises the question of how to build better tempering paths – that go beyond simple powering – to obtain fast mixing kernels. There are some recent developments on this issue in the context of PT/ST (Syed et al. (2021)). How to leverage these ideas while retaining the initial computational efficiency of cyclical MCMC is also an important direction for future research.

## 6 Proof of Theorem 7

For $j \ge 1$, let $\nu^{(j)}$ denote the marginal distribution of $\theta^{(j)}$, and define the kernel $\bar{M}_j$ by

$$\bar{M}_j(u, A) \stackrel{\text{def}}{=} \int_A M_j(u, \mathrm{d}v)\frac{\Pi_{j-1}(v)}{\Pi_j(v)}.$$

First, we observe that if $\nu^{(j)}$ admits a density with respect to $\Pi_j$, and $\mathrm{d}\nu^{(j)}/\mathrm{d}\Pi_j = f_j$, then $\nu^{(j+1)}$ admits

a density with respect to $\Pi_{j+1}$, and

$$\frac{\mathrm{d}\nu^{(j+1)}}{\mathrm{d}\Pi_{j+1}}(u) = \bar{M}_{j+1} f_j(u), \quad u \in \Theta.$$

To see this, use reversibility and write for any $A \in \mathcal{B}_\Theta$,

$$\begin{aligned}\nu^{(j+1)}(A) &= \int_\Theta \Pi_j(\mathrm{d}u) f_j(u) \int_\Theta M_{j+1}(u, \mathrm{d}v) \mathbf{1}_A(v) \\ &= \int_A \Pi_{j+1}(\mathrm{d}u) \int_\Theta M_{j+1}(u, \mathrm{d}v) f_j(v) \frac{\Pi_j(v)}{\Pi_{j+1}(v)}.\end{aligned}$$

Since, by assumption the initial distribution of the cyclical MCMC sampler has a density denoted $f_0$ with respect to $\Pi$, and $\nu^{(0)}(\mathrm{d}x) = f_0(x)\Pi_0(\mathrm{d}x)$, we conclude that for all $j \geq 1$, $\nu^{(j)}$ has a density $f_j$ with respect to $\Pi_j$, and the sequence $\{f_j,\ j \geq 0\}$ satisfies

$$f_{j+1} = \bar{M}_{j+1} f_j. \tag{19}$$

Using this and the Cauchy-Schwarz inequality, we write for all $j \geq 0$,

$$\|\nu^{(j)} - \Pi_j\|_{\mathrm{tv}} = \int_\Theta \left| \frac{\mathrm{d}\nu^{(j)}}{\mathrm{d}\Pi_j}(u) - 1 \right| \Pi_j(\mathrm{d}u) \leq \sqrt{\mathsf{Var}_j\left(f_j\right)}. \tag{20}$$

We show in Lemma 16 in the supplement that if $f \in \mathcal{L}^2_{j-1}$, then $\bar{M}_j f \in \mathcal{L}^2_j$, and the adjoint of the operator $\bar{M}_j$ is $M_j$. Also, recall that $f_j$ is the density of $\nu^{(j)}$ with respect to $\Pi_j$, and we have seen above that $f_j = \bar{M}_j f_{j-1}$, and $\Pi_j(\bar{M}_j f_{j-1}) = \Pi_{j-1}(f_{j-1}) = 1$. Therefore,

$$\begin{aligned}\mathsf{Var}_j(f_j) &= \int_\Theta (\bar{M}_j f_{j-1})^2 \Pi_j(\mathrm{d}x) - 1 \\ &= \left\langle \bar{M}_j f_{j-1}, \bar{M}_j f_{j-1} \right\rangle_j - 1 \\ &= \left\langle f_{j-1}, M_j \bar{M}_j f_{j-1} \right\rangle_{j-1} - 1 = \left\langle f_{j-1}, Q_j f_{j-1} \right\rangle_{j-1} - 1,\end{aligned}$$

where the operator $Q_j$ is as introduced in (10). Whereas

$$\mathsf{Var}_{j-1}(f_{j-1}) = \left\langle f_{j-1}, f_{j-1} \right\rangle_{j-1} - 1.$$

Hence

$$\begin{aligned}&\mathsf{Var}_j(f_j) \\ &\quad = \mathsf{Var}_{j-1}(f_{j-1}) - \left\langle f_{j-1}, (\mathbb{I} - Q_j) f_{j-1} \right\rangle_{j-1},\end{aligned} \tag{21}$$

where $\mathbb{I}$ denotes the identity operator. We now relate the term $\left\langle f_{j-1}, (\mathbb{I} - Q_j) f_{j-1} \right\rangle_{j-1}$ to the spectral gap $\lambda_j$ of $Q_j$. For $f \in \mathcal{L}^2_{j-1}$,

$$\begin{aligned}&\int_\Theta \int_\Theta \left(f(y) - f(x)\right)^2 \Pi_{j-1}(\mathrm{d}x) Q_j(x, \mathrm{d}y) \\ &= \int_\Theta \Pi_{j-1}(\mathrm{d}x) Q_j f^2(x) + \int_\Theta \Pi_{j-1}(\mathrm{d}x) f^2(x) Q_j(x, \Theta) \\ &\qquad - 2 \left\langle f, Q_j f \right\rangle_{j-1}.\end{aligned}$$

We have

$$\begin{aligned}&Q_j(x, \Theta) = 1 \\ &\quad + \int_\Theta M_j(x, \mathrm{d}y) \int_\Theta M_j(y, \mathrm{d}z) \left( \frac{\Pi_{j-1}(z)}{\Pi_j(z)} - 1 \right).\end{aligned}$$

Hence

$$\begin{aligned}&\int_\Theta \Pi_{j-1}(\mathrm{d}x) f^2(x) Q_j(x, \Theta) \\ &\qquad = \Pi_{j-1}(f^2) + \int_\Theta \Pi_{j-1}(\mathrm{d}x) f^2(x) \\ &\quad \times \int_\Theta M_j(x, \mathrm{d}y) \int_\Theta M_j(y, \mathrm{d}z) \left( \frac{\Pi_{j-1}(z)}{\Pi_j(z)} - 1 \right).\end{aligned}$$

And since $\Pi_j(Q_j f) = \Pi_{j-1}(f)$, we have

$$\begin{aligned}&\int_\Theta \Pi_{j-1}(\mathrm{d}x) Q_j f^2(x) \\ &= \Pi_{j-1}(f^2) + \int_\Theta \Pi_j(\mathrm{d}x) \left( \frac{\Pi_{j-1}(x)}{\Pi_j(x)} - 1 \right) Q_j f^2(x) \\ &\qquad = \Pi_{j-1}(f^2) + \int_\Theta \Pi_{j-1}(\mathrm{d}x) f^2(x) \\ &\quad \times \int_\Theta M_j(x, \mathrm{d}y) \int_\Theta M_j(y, \mathrm{d}z) \left( \frac{\Pi_{j-1}(z)}{\Pi_j(z)} - 1 \right).\end{aligned}$$

We conclude that

$$\begin{aligned}&\frac{1}{2} \int_\Theta \int_\Theta \left(f(y) - f(x)\right)^2 \Pi_{j-1}(\mathrm{d}x) Q_j(x, \mathrm{d}y) \\ &\qquad = \left\langle f, (\mathbb{I} - Q_j) f \right\rangle_{j-1} \\ &\qquad + \int_\Theta \Pi_{j-1}(\mathrm{d}x) f^2(x) \\ &\quad \times \int_\Theta M_j(x, \mathrm{d}y) \int_\Theta M_j(y, \mathrm{d}z) \left( \frac{\Pi_{j-1}(z)}{\Pi_j(z)} - 1 \right).\end{aligned} \tag{22}$$

Since $Q_j$ is positive as shown in Lemma 16, with $\bar{f} = f - \Pi_{j-1}(f)$, we have $\left\langle \bar{f}, (\mathbb{I} - Q_j) \bar{f} \right\rangle_{j-1} \leq \mathsf{Var}_{j-1}(f)$, so that

$$\begin{aligned}&\frac{1}{2} \int_\Theta \int_\Theta \left(f(y) - f(x)\right)^2 \Pi_{j-1}(\mathrm{d}x) Q_j(x, \mathrm{d}y) \\ &\qquad \leq (1 + \alpha_j) \mathsf{Var}_{j-1}(f),\end{aligned}$$

which implies

$$\lambda_j \leq 1 + \alpha_j. \tag{23}$$

Applied to $f_{j-1}$, (22) gives

$$\begin{aligned}&- \left\langle f_{j-1}, (\mathbb{I} - Q_j) f_{j-1} \right\rangle_{j-1} \\ &\leq -\frac{1}{2} \int_\Theta \int_\Theta \left(f_{j-1}(y) - f_{j-1}(x)\right)^2 \Pi_{j-1}(\mathrm{d}x) Q_j(x, \mathrm{d}y) \\ &\qquad + \alpha_j \int_\Theta \Pi_{j-1}(\mathrm{d}x) f_{j-1}^2(x) \\ &\leq -\lambda_j \mathsf{Var}_{j-1}(f_{j-1}) + \alpha_j \mathsf{Var}_{j-1}(f_{j-1}) + \alpha_j.\end{aligned}$$

Taking this last display in (21), we deduce that

$$\mathsf{Var}_j(f_j) \leq (1-\lambda_j+\alpha_j)\,\mathsf{Var}_{j-1}(f_{j-1}) + \alpha_j.$$

Iterating this inequality yields

$$\mathsf{Var}_j(f_j) \leq \mathsf{Var}_0(f_0)\prod_{k=1}^{j}(1-\lambda_k+\alpha_k) + \sum_{i=1}^{j}\alpha_i \prod_{\ell=i+1}^{j}(1-\lambda_\ell+\alpha_\ell).$$

The result follows from the last display and (20).

## 7 Proof of Theorem 13

Fix a measurable set $A$ and starting point $\theta$. Since $\nu^{(L)} = \nu^{(L_2)}\prod_{i=L_2+1}^{L} M_i$, we have

$$\nu^{(L)}(A) \geq \sum_{j=1}^{d}\nu^{(L_2)}(I_j)\inf_{\theta'\in I_j}\Big(\prod_{i=L_2+1}^{L} M_{i,n}\Big)(\theta', A) \geq \sum_{j=1}^{d}\nu^{(L_2)}(I_j)(\Pi^{(j)}(A)-\delta),$$

using Assumptions 2 and 3. Similarly,

$$\begin{aligned}\nu^{(L)}(A) &\leq \sum_{j=1}^{d}\nu^{(L_2)}(I_j) \\ &\quad\times \sup_{\theta'\in I_j}\Big(\prod_{i=L_2}^{L} M_{i,n}\Big)(\theta', A) + \Big(1-\sum_{j=1}^{d}\nu^{(L_2)}(I_j)\Big) \\ &\leq \sum_{j=1}^{d}\nu^{(L_2)}(I_j)(\Pi^{(j)}(A)+\delta) + \Big(1-\sum_{j=1}^{d}\nu^{(L_2)}(I_j)\Big).\end{aligned}$$

The proof follows from combining these two inequalities.

## 8 Acknowledgments

This work is partially funded by the NSF grant DMS-2210664. We are also grateful to Joonha Park for very helpful comments on an initial draft of this work.

## References

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2016). Variational Inference: A Review for Statisticians. *ArXiv e-prints* arXiv:1601.00670.

Chen, T., Fox, E. B. and Guestrin, C. (2014). Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14, JMLR.org.

Douc, R., Moulines, É., Priouret, P. and Soulier, P. (2018). *Markov chains*. Springer International Publishing.

Durmus, A., Moulines, E. et al. (2019). High-dimensional bayesian inference via the unadjusted langevin algorithm. *Bernoulli* **25** 2854–2882.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*. PMLR.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185.

Geyer, C. (1991). Markov chain monte carlo maximum likelihood, in computing science and statistics: Proceedings of the 32rd symposium on the interface, ed. e.m. keramigas, fairfax: Interface foundation, pp 156-163 .

Geyer, C. J. and Thompson, E. (1995). Annealing markov chain monte carlo with applications to pedigree analysis. *Journal of the American Statistical Association* **90** 909–920.

Graves, A. (2011). Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems* (J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger, eds.), vol. 24. Curran Associates, Inc.

Hervé, L. (2008). Vitesse de convergence dans le théorème limite central pour des chaînes de Markov fortement ergodiques. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** 280–292.

Hukushima, K. and Nemoto, K. (1996). Exchange monte carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan* **65** 1604–1608.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* **110** 457–506.

Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of metropolis algorithms. *Sto. Proc. Appl.* **85** 341–361.

Kontoyiannis, I., Lastras-Montao, L. A. and Meyn, S. P. (2006). Exponential bounds and stopping rules for mcmc and general markov chains. *First International Conference on Performance Evaluation Methodologies and Tools, Pisa, Italy* .

Ma, Y.-A., Chen, T. and Fox, E. B. (2015). A complete recipe for stochastic gradient mcmc. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'15, MIT Press, Cambridge, MA, USA.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new monte carlo schemes. *Europhysics letters* **19** 451–458.

Matthews, P. (1993). A slowly mixing markov chain with implications for gibbs sampling. *Statistics & Probability Letters* **17** 231–236.

Meyn, S. and Tweedie, R. L. (2009). *Markov chains and stochastic stability*. 2nd ed. Cambridge University Press, Cambridge.

Pittorino, F., Lucibello, C., Feinauer, C., Perugini, G., Baldassi, C., Demyanenko, E. and Zecchina, R. (2021). Entropic gradient descent algorithms and wide flat minima. In *International Conference on Learning Representations*.
URL `https://openreview.net/forum?id=xjXg0bnoDmS`

Raginsky, M., Rakhlin, A. and Telgarsky, M. (2017). Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory* (S. Kale and O. Shamir, eds.), vol. 65 of *Proceedings of Machine Learning Research*. PMLR.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. 2nd ed. Springer Texts in Statistics, Springer-Verlag, New York.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15** 1929–1958.

Syed, S., Romaniello, V., Campbell, T. and Bouchard-Cote, A. (2021). Parallel tempering on optimized paths. In *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*. PMLR.
URL `https://proceedings.mlr.press/v139/syed21a.html`

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11, Omnipress, USA.

Whiteley, N. (2013). Stability properties of some particle filters. *The Annals of Applied Probability* **23** 2500 – 2537.

Woodard, D., Schmidler, S. and Huber, M. (2009a). Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electron. J. Probab.* **14** 780–804.
URL `https://doi.org/10.1214/EJP.v14-638`

Woodard, D. B., Schmidler, S. C. and Huber, M. (2009b). Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Probab.* **19** 617–640.

Zhang, R., Li, C., Zhang, J., Chen, C. and Wilson, A. G. (2019). Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932* .

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Not Applicable]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# On cyclical MCMC sampling: Supplementary Materials

## 9 PROOF OF PROPOSITION 4

In what follows we write $\beta_j$ as a short for $\beta(j/L)$. We have

$$\frac{\Pi_{j-1}(z)}{\Pi_j(z)} = \exp\left((\beta_j - \beta_{j-1})\mathcal{E}(z) - \log(Z_{j-1}/Z_j)\right).$$

For $t \in [0,1]$, let $\Pi_{j,t}(\theta) \propto \exp(-(t\beta_j + (1-t)\beta_{j-1})\mathcal{E}(\theta))$ be a probability measure that interpolates between $\Pi_{j-1}$ and $\Pi_j$. By the path sampling identity (Gelman and Meng (1998)),

$$\log\left(\frac{Z_{j-1}}{Z_j}\right) = -(\beta_j - \beta_{j-1})\int_0^1 \int_\Theta \mathcal{E}(u)\Pi_{j,t}(u)\mathrm{d}u.$$

We deduce that

$$(\beta_j - \beta_{j-1})\mathcal{E}(z) - \log\left(\frac{Z_{j-1}}{Z_j}\right) = (\beta_j - \beta_{j-1})\int_0^1 \int_\Theta (\mathcal{E}(u) - \mathcal{E}(z))\Pi_{j,t}(u)\mathrm{d}u.$$

Since $|e^x - 1| \leq xe^{|x|}$, we obtain

$$\left|\frac{\Pi_{j-1}(z)}{\Pi_j(z)} - 1\right| \leq |\beta_j - \beta_{j-1}|\mathsf{osc}(\mathcal{E})e^{|\beta_j - \beta_{j-1}|\mathsf{osc}(\mathcal{E})}.$$

We take $C = \mathsf{osc}(\mathcal{E}) \times \max_{1\leq j\leq L} e^{|\beta_j - \beta_{j-1}|\mathsf{osc}(\mathcal{E})}$, and the result follows by first order Taylor expansion of $\beta$.

## 10 ON THE TERM $\mathsf{Var}_0(f_0)$ IN THEOREM 7

One natural objection to the conclusion of Theorem 7 is: for typical MCMC algorithms, the variance term $\mathsf{Var}_0(f_0)$ appearing in the conclusion of Theorem 7 is infinite, and so the theorem does not give nontrivial upper bounds. Fortunately, in many examples, this problem can be easily fixed via truncation. More precisely, assume there exist constants $0 < \omega < 1$, $0 < C < \infty$ and $T \in \mathbb{N}$ so that $\prod_{t=0}^T M_t$ can be written in the form

$$\prod_{t=0}^T M_t(\theta, \cdot) = \omega H(\cdot) + (1-\omega)R_T(\theta, \cdot),$$

where $H$ has a density $h$ satisfying $\mathsf{Var}_T(h) \leq C$. In this case, one has

$$\|\prod_{s=0}^L M_s(\theta, \cdot) - \Pi\|_{\mathrm{tv}} \leq \|H \prod_{s=T+1}^L M_s - \Pi\|_{\mathrm{tv}} + 2(1-\omega), \tag{24}$$

and so Theorem 7 can be applied to the first term starting from time $T$, with the variance term bounded by $C$.

## 11 ON THE KERNELS $Q_j$'s

**Lemma 16.** *Assume H1. Then for $f, h \in \mathcal{L}^2_{j-1}$, and $g \in \mathcal{L}^2_j$, $\bar{M}_j f \in \mathcal{L}^2_j$, $\left\langle \bar{M}_j f, g \right\rangle_j = \langle f, M_j g \rangle_{j-1}$. Furthermore, $\langle f, Q_j h \rangle_{j-1} = \langle h, Q_j f \rangle_{j-1}$, and $\langle f, Q_j f \rangle_{j-1} \geq 0$.*

*Proof.* Set $\rho_j \stackrel{\text{def}}{=} \|\Pi_{j-1}/\Pi_j\|_\infty$. For $f \in \mathcal{L}^2_{j-1}$, the fact that $\bar{M}_j f$ belongs to $\mathcal{L}^2_j$ follows from the Cauchy-Schwarz inequality and $\Pi_j M_j = \Pi_j$. Indeed,

$$\int_\Theta \Pi_j(\mathrm{d}u)|\bar{M}_j f(u)|^2 \leq \int_\Theta \Pi_j(\mathrm{d}u) \int_\Theta M_j(u, \mathrm{d}v) f^2(v) \left|\frac{\Pi_{j-1}(v)}{\Pi_j(v)}\right|^2 = \int_\Theta f^2(v) \frac{\Pi_{j-1}(v)}{\Pi_j(v)} \Pi_{j-1}(\mathrm{d}v) \leq \rho_j \Pi_{j-1}(f^2).$$

The equality $\left\langle \bar{M}_j f, g \right\rangle_j = \langle f, M_j g \rangle_{j-1}$ follows easily from reversibility. Indeed, since $\bar{M}_j f \in \mathcal{L}^2_j$, we have

$$\begin{aligned}\left\langle \bar{M}_j f, g \right\rangle_j &= \int_\Theta \Pi_j(\mathrm{d}u) \int_\Theta M_j(u, \mathrm{d}v) g(u) f(v) \frac{\Pi_{j-1}(v)}{\Pi_j(v)} \\ &= \int_\Theta \Pi_j(\mathrm{d}u) \int_\Theta M_j(u, \mathrm{d}v) g(v) f(u) \frac{\Pi_{j-1}(u)}{\Pi_j(u)} = \langle f, M_j g \rangle_{j-1}.\end{aligned}$$

A similar argument as above shows that for $f \in \mathcal{L}^2_{j-1}$,

$$\int_\Theta \left(Q_j f(x)\right)^2 \Pi_{j-1}(\mathrm{d}x) \leq \rho_j^2 \Pi_{j-1}(f^2),$$

and

$$\langle f, Q_j h \rangle_{j-1} = \int_\Theta \Pi_j(\mathrm{d}u) \int_\Theta M_j^2(u, \mathrm{d}v) f(u) h(v) \frac{\Pi_{j-1}(u)}{\Pi_j(u)} \frac{\Pi_{j-1}(v)}{\Pi_j(v)} = \langle h, Q_j f \rangle_{j-1}.$$

The positivity is easily seen by noting that $Q_j$ is the product of $M_j$ and its adjoint $\bar{M}_j$. □

## 12 CHECKING ASSUMPTION H2

We first check that a Lyapunov drift condition implies Assumption 2. Specifically, suppose that there exists a function $V : \Theta_j \to [1, \infty)$ and constants $m > 0$, $0 < \alpha \leq 1$, and $0 \leq \beta < \infty$ with following properties:

1. For $\theta \notin \Theta_j$, we have $V(\theta) \geq e^m$.
2. For all $\theta \in \Theta_j$ and all $L_2 \leq i \leq L$, one has

$$M_i V(\theta) \leq (1 - \alpha) V(\theta) + \beta.$$

Then we can calculate that for all $\theta \in I_1$ and all $L_2 \leq i \leq L$ we have

$$\mathbb{E}_\theta \left(V(\theta^{(i)})\right) \leq \frac{\beta}{\alpha}.$$

Thus, by Markov's inequality,

$$\mathbb{P}\left(X_i \notin \Theta_j\right) \leq \mathbb{P}\left(V(X_i) > e^m\right) \leq e^{-m} \mathbb{E}\left(V(X_i)\right) \leq \frac{\beta e^{-m}}{\alpha}.$$

Taking a union bound over $L_2 \leq i \leq L$, we find that Assumption 2 holds with exponentially small constant $\delta_1$ under some mild assumption on the level $m$ of the drift function $V$.

We next note that this proof approach can be used to verify Assumption 2 for targets given in Equation (13) of the main document for sufficiently small $\sigma$. Toward that we take sets $\Theta_1 = [-1.5, -0.5]$ and $\Theta_2 = [0.5, 1.5]$ and time $\frac{L}{3} \leq L_2 \leq \frac{2L}{3}$. For $\sigma, s, c > 0$, we consider the Metropolis-Hastings chain $K = K_{\sigma,s,c}$ with target $\mathcal{N}(0, sc^2\sigma^2)$ and proposal $Q(\theta, \cdot) = Unif([\theta - s\sigma, \theta + s\sigma])$. Denote by $a_n$ the acceptance function associated with this Metropolis-Hastings kernel. Finally, for $\alpha > 0$ we consider candidate Lyapunov functions of the form $V(\theta) = e^{\frac{\alpha}{\sigma}|\theta|}$.

**Lemma 17.** *There exists $\epsilon > 0$ so that for all $s \in [-0.5, 2]$ and all $0 < \sigma < \alpha < \epsilon$ sufficiently small, $V$ is a Lyapunov function satisfying:*

$$(KV)(\theta) \leq 0.7V(\theta) + e^{2s\alpha}\mathbf{1}_{|\theta|\leq s\sigma}.$$

**Remark 18.** Although we state the result for **exactly** Gaussian targets on the **full** real line, it applies with no substantial changes in the following situations:

1. **Restricted State Space:** If we change the target from the Gaussian $\mathcal{N}(0, sc^2\sigma^2)$ to the restriction of this Gaussian to an interval $[A, B]$, the result applies with no changes to the constants as long as $A < -s\sigma$ and $B > s\sigma$. To verify this, note that rejecting proposals from within $[A, B]$ to points outside of $[A, B]$ can only decrease the value of $V$.
2. **Small Multiplicative Perturbations:** Denote by $f$ the density of $\mathcal{N}(0, sc^2\sigma^2)$. If we target instead a distribution with unimodal density $g$ satisfying $\left|\frac{g(\theta)}{f(\theta)} - 1\right| \leq \epsilon V(\theta)$, then the same Lyapunov inequality holds with the constant 0.7 replaced by $(0.7 + \epsilon)$.

These observations can be combined to verify Lyapunov conditions for Gaussian mixture models restricted to regions where one mixture component dominates the density.

*Proof.* Let $\theta \in \mathbb{R}$ and let $\Theta_1 \sim K(\theta, \cdot)$. We consider three cases: $\theta < -s\sigma$, $\theta > s\sigma$, or $\theta \in [-s\sigma, s\sigma]$.

In the first case,

$$\begin{aligned}
&\frac{2}{s\sigma}\mathbb{E}[V(\Theta_1)] \\
&\quad= \int_{-s\sigma}^{0} V(\theta+x)a_n(\theta,\theta+x)dx + \int_{-s\sigma}^{0} V(\theta)(1-a_n(\theta,\theta+x))dx + \int_{0}^{s\sigma} V(\theta+x)dx \\
&\quad= V(\theta)\left(\int_{-s\sigma}^{0}(V(x)-1)a_n(\theta,\theta+x)dx + \frac{1}{s\sigma} + \int_{0}^{s\sigma} V(x)dx\right) \\
&\quad= V(\theta)\left(\int_{-s\sigma}^{0}(V(x)-1)a_n(\theta,\theta+x)dx + \frac{1}{s\sigma} + \frac{\sigma}{\alpha}(1-e^{-\alpha s})\right).
\end{aligned}$$

For fixed $s$ and all $0 < \sigma < \alpha < 0.1$, we can bound the last term and continue:

$$\begin{aligned}
&\frac{2}{s\sigma}\mathbb{E}[V(\Theta_1)] \\
&\quad\leq V(\theta)\left(\frac{1.1}{s\sigma} + \int_{-s\sigma}^{0}(V(x)-1)a_n(\theta,\theta+x)dx\right) \\
&\quad= V(\theta)\left(\frac{1.1}{s\sigma} + \int_{-s\sigma}^{0}(e^{\frac{\alpha}{\sigma}|x|}-1)e^{\frac{1}{2sc^2\sigma^2}(\theta^2-(\theta+x)^2)}dx\right) \\
&\quad= V(\theta)\left(\frac{1.1}{s\sigma} + \int_{-s\sigma}^{0}(e^{\frac{\alpha}{\sigma}|x|}-1)e^{-\frac{1}{2sc^2\sigma^2}(2\theta x+x^2)}dx\right) \\
&\quad\leq V(\theta)\left(\frac{1.1}{s\sigma} + \int_{-s\sigma}^{0} e^{-\frac{\alpha}{\sigma}x-\frac{1}{2sc^2\sigma^2}(2\theta x+x^2)}dx\right).
\end{aligned}$$

Investigating the two terms appearing inside the exponential, we note that for small $\alpha$, the second term dominates. Thus, for sufficiently small $0 < \sigma < \alpha < \epsilon$, we have

$$\frac{2}{s\sigma}\mathbb{E}[V(\Theta_1)] \leq V(\theta)\frac{1.2}{s\sigma}.$$

This proves the desired inequality in the first case.

In the second case, we simply note that $|\Theta_1| \leq |\theta| + s\sigma$, so

$$V(\Theta_1) \leq V(2s\sigma) = e^{2\alpha s} \leq 0.7\, V(\theta) + e^{2\alpha s}.$$

This proves the desired inequality in the second case.

The third case is essentially identical to the first case. □