Stability of nonlinear flows of probability measures related to sequential Markov Chain Monte Carlo methods

Andreas Eberle Carlo Marinelli

April 26, 2011

Abstract

Sequential Monte Carlo Samplers are a class of stochastic algorithms for Monte Carlo integral estimation w.r.t. probability distributions which combine elements of Markov chain Monte Carlo methods and importance sampling/resampling schemes. We develop a stability analysis by functional inequalities for a nonlinear flow of probability measures describing the limit behavior of the algorithms as the number of particles tends to infinity. Stability results are derived both under global and local assumptions on the generator of the underlying Metropolis dynamics. This allows us to prove that the combined methods sometimes have good asymptotic stability properties in multimodal setups where traditional MCMC methods mix extremely slowly. For example, this holds for the mean field Ising model at all temperatures.

1 Introduction

Spectral gap estimates and related functional inequalities provide powerful tools for the study of convergence to equilibrium of reversible time-homogeneous Markov processes (see e.g. [2], [3], [4]). They have been employed successfully to analyze convergence properties of Markov Chain Monte Carlo (MCMC) methods for approximate sampling and integral estimation with respect to a fixed probability measure μ, see e.g. [12] and references therein. In several applications of MCMC methods one is further interested in estimating sequentially expectation values w.r.t. evolving probability measures $(\mu_t)_{t\geq 0}$. For example, the measures μ_t may be used as the basis of a homotopy method for interpolating between a nice initial distribution μ_0 and a target distribution μ_β that cannot be approximated directly in a feasible way (cf. the examples in XXX). Corresponding sequential MCMC methods rely on replacing the Kolmogorov forward equation of a stationary time-homogeneous Markov process by a nonlinear Fokker-Planck equation that is satisfied by the measures μ_t. The empirical distribution of an interacting particle system discretizing the Fokker-Planck equation are then used as estimators for expectation values w.r.t. μ_t, cf. [DMDJ], [arXiv '10]. In [arXiv '10] we apply functional inequalities to derive non-asymptotic error bounds for the particle system approximations. The results are partially restrictive, since the validity of appropriate logarithmic Sobolev inequalities is assumed. The purpose of the present article is to show that under less restrictive assumptions, functional inequalities can be used efficiently to study the stability properties of the limiting nonlinear Fokker-Planck equation.

2 Setup

2.1 Evolving probability measures

Let $(\mu_t)_{t\geq 0}$ denote a family of mutually absolutely continuous probability measures on a set S. To keep the presentation as simple and non-technical as possible, we assume that S is finite.

We assume that the measures are represented in the form

$$\mu_t(x) = \frac{1}{Z_t} \exp\left(-\mathcal{U}_t(x)\right)\, \mu_0(x), \qquad t \geq 0, \tag{1}$$

where Z_t is a normalization constant, and $(t, x) \mapsto \mathcal{U}_t(x)$ is a given function on $[0, \infty) \times S$ that is continuously differentiable in the first variable. If, for example, $\mathcal{U}_t(x) = t\mathcal{U}(x)$ for some function $\mathcal{U} : S \to \mathbb{R}$, then $(\mu_t)_{t\geq 0}$ is the exponential family corresponding to $\mathcal{U}$ and μ_0. Let

$$H_t(x) := -\frac{\partial}{\partial t} \log \mu_t(x) = -\frac{\partial}{\partial t} \log \frac{\mu_t(x)}{\mu_0(x)}$$

denote the negative logarithmic time derivative of the measures μ_t. Note that

$$\mu_t(x) = \exp\left(-\int_0^t H_s(x)\, ds\right)\, \mu_0(x)\,, \tag{2}$$

and

$$\langle H_t, \mu_t \rangle = -\frac{d}{dt}\mu_t(S) = 0 \qquad \text{for all } t \geq 0, \tag{3}$$

where

$$\langle f, \nu \rangle := \int_S f\, d\nu \ = \sum_{x\in S} f(x)\, \nu(x)$$

denotes the integral of a function $f : S \to \mathbb{R}$ w.r.t. a measure ν on S. In particular,

$$H_t = \frac{\partial}{\partial t}\mathcal{U}_t - \left\langle \frac{\partial}{\partial t}\mathcal{U}_t, \mu_t \right\rangle .$$

In applications we have in mind, the functions $\mathcal{U}_t$ are given explicitly. Hence H_t is known explicitly up to an additive time-dependent constant. The evaluation of this constant, however, would require integration w.r.t. μ_t.

If all the functions H_t, $t \geq 0$, vanish then $\mu_t = \mu_0$ for all $t \geq 0$. In this case the measures are invariant for a Markov transition semigroup $(p_t)_{t\geq 0}$, i.e.,

$$\mu_s p_{t-s} = \mu_t \qquad \text{for all } t \geq s \geq 0,$$

provided the generator $\mathcal{L}$ satisfies $\mu_0\mathcal{L} = 0$, i.e.

$$\sum_{x\in S} \mu_0(x)\mathcal{L}(x, y) = 0 \qquad \text{for all } y \in S.$$

This fact is exploited in Markov Chain Monte Carlo (MCMC) methods for approximating expectation values w.r.t. the measure μ_0. The particle systems studied below can be applied for the same purpose when the measures μ_t are time-dependent.

2.2 Associated Fokker-Planck equations

To obtain approximations of the measures μ_t, we consider generators (Q-matrices) $\mathcal{L}_t$, $t \geq 0$, of a time-inhomogeneous Markov process on S satisfying the detailed balance conditions

$$\mu_t(x)\mathcal{L}_t(x,y) = \mu_t(y)\mathcal{L}_t(y,x) \quad \forall\, t \geq 0,\ x,y \in S. \tag{4}$$

For example, $\mathcal{L}_t$ could be the generator of a Metropolis dynamics w.r.t. μ_t, i.e.

$$\mathcal{L}_t(x,y) \;=\; K_t(x,y) \cdot \min\left(\frac{\mu_t(y)}{\mu_t(x)}, 1\right) \text{ for } x \neq y,$$

$\mathcal{L}_t(x,x) = -\sum_{y \neq x} \mathcal{L}_t(x,y)$, where the proposal matrix K_t is a given symmetric transition matrix on S. In the sequel we will use the notation $\mathcal{L}_t^* \mu$ to denote the adjoint action of the generator on a probability measure, i.e.

$$(\mathcal{L}_t^*\mu)(y) := (\mu\mathcal{L}_t)(y) = \sum_{x \in S} \mu(x)\mathcal{L}_t(x,y).$$

By (4), $\mathcal{L}_t^*\mu_t = 0$, i.e.

$$\langle \mathcal{L}_t f, \mu_t \rangle = 0 \qquad \text{for all } f : S \to \mathbb{R} \text{ and } t \geq 0.$$

We fix non-negative constants λ_t, $t \geq 0$. Since the state space S is finite, the measures μ_t are the unique solution of the evolution equation for measures

$$\frac{\partial}{\partial t}\nu_t = \lambda_t\, \mathcal{L}_t^* \nu_t - H_t \nu_t \tag{5}$$

with initial condition $\nu_0 = \mu_0$. In general, solutions of (5) are not necessarily probability measures, even if ν_0 is a probability measure. Therefore, we consider the equation

$$\frac{\partial}{\partial t}\eta_t \;=\; \lambda_t\, \mathcal{L}_t^*\eta_t \;-\; H_t\eta_t \;+\; \langle H_t, \eta_t \rangle\; \eta_t \tag{6}$$

satisfied by the normalized measures $\eta_t = \frac{\nu_t}{\nu_t(S)}$. Note that, by (3), μ_t also solves (6). Moreover, if η_t is a solution of (6), then

$$\nu_t = \exp\left(-\int_0^t \langle H_s, \eta_s \rangle\right) \eta_t$$

is the unique solution of (5) with initial condition $\nu_0 = \eta_0$.

The Fokker-Planck equation (6) is an evolution equation for probability measures which, in contrast to the unnormalized equation, is not modified by adding constants to the functions H_t.

2.3 Stability

Let η_t be the solution of (6) with initial condition $\nu \in \mathcal{M}_1(S)$, and let

$$g_t(y) := \frac{\eta_t(y)}{\mu_t(y)}, \qquad t \geq 0,$$

denote the relative density of η_t w.r.t. the measure μ_t defined by (1). Moreover, let

$$\varepsilon_t := \mathbb{E}_t[(g_t - 1)^2] = \langle (g_t - 1)^2, \mu_t \rangle$$

denote the mean square error (χ^2–contrast) of η_t w.r.t. μ_t.

The main objective of this article is to develop efficient tools to bound the growth of ε_t. The analysis will be based on Theorem 1. To estimate the right-hand side of (8) we have to control the two terms involving H_t (which correspond to importance sampling/resampling) by the Dirichlet form $\mathcal{E}_t$ (which corresponds to MCMC moves). We first discuss how this can be achieved in the presence of a good global spectral gap estimate. Afterwards, we give results based on local Poincaré-type inequalities, which can sometimes be used to control the error growth in multimodal setups where good global mixing properties of the underlying Markov chains do not hold.

3 Stability based on global estimates

3.1 Preliminaries

The associated Dirichlet form on functions $f, g : S \to \mathbb{R}$ is

$$\mathcal{E}_t(f,g) := -\mathbb{E}_t[f\,\mathcal{L}_t g] = \frac{1}{2} \sum_{x,y \in S} (f(y) - f(x))(g(y) - g(x))\,\mathcal{L}_t(x,y)\,\mu_t(x),$$

where $\mathbb{E}_t$ stands for expectation w.r.t. μ_t, and

$$(\mathcal{L}_t g)(x) := \sum_y \mathcal{L}_t(x,y) g(y).$$

We shall often use the abbreviated notation $\mathcal{E}_t(f) := \mathcal{E}_t(f,f)$.

Now we have a first result about time evolution of the mean square error.

Theorem 1. *The densities g_t solve*

$$\frac{\partial}{\partial t} g_t = \lambda_t \mathcal{L}_t g_t + \mathbb{E}_t[H_t g_t]\, g_t. \tag{7}$$

Moreover, the time evolution of the mean square error is given by

$$\frac{1}{2}\frac{d}{dt}\varepsilon_t = -\lambda_t\, \mathcal{E}_t(g_t - 1) - \frac{1}{2}\mathbb{E}_t[H_t(g_t-1)^2] + \mathbb{E}_t[H_t(g_t - 1)]\,\varepsilon_t. \tag{8}$$

Proof. To simplify the notation, we assume $\lambda_t = 1$ for all $t \geq 0$. The general case is similar with $\mathcal{L}_t$ replaced by $\lambda_t \mathcal{L}_t$. Let us first derive equation (7): since μ_t has full support and is differentiable in t, we have

$$\frac{\partial}{\partial t} g_t = \frac{1}{\mu_t}\frac{\partial}{\partial t}\eta_t - \frac{\eta_t}{\mu_t}\frac{\partial}{\partial t}\log \mu_t. \tag{9}$$

Note that, by the detailed balance condition (4), the relative density of $\mathcal{L}_t^* \eta_t$ w.r.t. μ_t is

$$\frac{\mathcal{L}_t^*\eta_t(y)}{\mu_t(y)} = \sum_x \eta_t(x)\,\frac{\mathcal{L}_t(x,y)}{\mu_t(y)} = \sum_x \eta_t(x)\,\frac{\mathcal{L}_t(y,x)}{\mu_t(x)} = \mathcal{L}_t g_t(y).$$

Hence (6) yields

$$\begin{aligned}\frac{1}{\mu_t}\frac{\partial}{\partial t}\eta_t &= \mathcal{L}_t g_t - H_t g_t + \langle H_t, \eta_t\rangle \, g_t \\ &= (\mathcal{L}_t - H_t) g_t + \langle H_t g_t, \mu_t\rangle \, g_t. \end{aligned} \tag{10}$$

Recalling that $H_t(x) := -\frac{\partial}{\partial t}\log \mu_t(x)$, $x \in S$, one has

$$-\frac{\eta_t}{\mu_t}\frac{\partial}{\partial t}\log \mu_t = H_t g_t. \tag{11}$$

Inserting (10) and (11) into (9) we obtain (7).

Let us now derive the equation for the quadratic error

$$\varepsilon_t = \left\langle (g_t - 1)^2, \mu_t \right\rangle \equiv \mathbb{E}_t\big[(g_t - 1)^2\big].$$

Differentiating this expression with respect to t one gets

$$\frac{d}{dt}\varepsilon_t = 2\left\langle (g_t - 1)\partial_t g_t, \mu_t\right\rangle + \left\langle (g_t - 1)^2, \partial_t \mu_t\right\rangle,$$

where, by (6),

$$\begin{aligned}\langle (g_t - 1)\partial_t g_t, \mu_t\rangle &= \big\langle (g_t - 1)\big(\mathcal{L}_t g_t + \langle H_t g_t, \mu_t\rangle \, g_t\big), \mu_t\big\rangle \\ &= \langle \mathcal{L}_t (g_t - 1), (g_t - 1)\rangle + \langle H_t g_t, \mu_t\rangle \langle g_t(g_t - 1), \mu_t\rangle \\ &= -\mathcal{E}_t(g_t - 1) + \langle H_t (g_t - 1), \mu_t\rangle \, \varepsilon_t, \end{aligned}$$

and, taking again into account that $H_t := -\frac{\partial}{\partial t}\log \mu_t$,

$$\left\langle (g_t - 1)^2, \partial_t \mu_t\right\rangle = \left\langle (g_t - 1)^2, (\partial_t \log \mu_t)\mu_t\right\rangle = -\left\langle H_t (g_t - 1)^2, \mu_t\right\rangle.$$

In the above derivation we have used the identities $\langle g_t - 1, \mu_t\rangle = \eta_t(S) - \mu_t(S) = 0$, $\mathcal{L}_t 1 = 0$, and $\langle H_t, \mu_t\rangle = 0$. We have thus proved (8) in the case $\lambda_t \equiv 1$. The general case follows similarly. □

For $t \geq 0$ let

$$C_t := \sup\left\{ \mathbb{E}_t[f^2]/\mathcal{E}_t(f) \,\middle|\, f : S \to \mathbb{R} \text{ s.t. } \mathbb{E}_t[f] = 0\,,\ f \not\equiv 0 \right\}$$

denote the (possibly infinite) inverse spectral gap of $\mathcal{L}_t$, and let

$$A_t := \sup\left\{ \mathbb{E}_t[H_t^- f^2]/\mathcal{E}_t(f) \,\middle|\, f : S \to \mathbb{R} \text{ s.t. } \mathbb{E}_t[f] = 0\,,\ f \not\equiv 0 \right\}.$$

Thus C_t and A_t are the optimal constants in the global Poincaré inequalities

$$\mathrm{Var}_t(f) \leq C_t\, \mathcal{E}_t(f) \qquad \forall f : S \to \mathbb{R}, \tag{12}$$

$$\mathbb{E}_t[H_t^-(f - \mathbb{E}_t[f])^2] \leq A_t\, \mathcal{E}_t(f) \qquad \forall f : S \to \mathbb{R}. \tag{13}$$

Here Var_t stands for variance w.r.t. μ_t.

Remark 2. (i) There exist efficient techniques to obtain upper bounds for C_t, for example the method of canonical paths, comparison methods (see e.g. [10]), as well as decomposition methods (see e.g. [6]). Variants of these techniques can be applied to estimate A_t as well.

(ii) Clearly, one has

$$A_t \;\leq\; C_t \cdot \sup_{x\in S} H_t^-(x), \tag{14}$$

so an upper bound on C_t yields a trivial (and usually far from optimal) upper bound on A_t.

Let

$$\sigma_t(H) \;:=\; \mathrm{Var}_t(H)^{1/2} \;=\; \mathbb{E}_t[H_t^2]^{1/2}$$

denote the standard deviation of H w.r.t. μ_t. The next result bounds the error growth in terms of C_t and A_t.

Theorem 3. *If $\lambda_t \geq A_t/2$ for all $t \geq 0$, then*

$$\frac{d}{dt}\log\varepsilon_t \leq -\frac{2\lambda_t - A_t}{C_t} \,+\, 2\,\sigma_t(H)\varepsilon_t^{1/2} \tag{15}$$

and

$$\frac{d}{dt}\log\varepsilon_t \leq -\frac{2\lambda_t - A_t}{C_t} + 2\Big(\frac{A_t}{C_t}\mathbb{E}_t[H_t^-]\Big)^{1/2}\varepsilon_t^{1/2} + \mathbb{E}_t[H_t^-]\,\varepsilon_t. \tag{16}$$

Inequality (15) is straightforward to prove, but sometimes (16) is stronger, since the constants only depend on the negative part of H_t.

Proof. We have to estimate the terms on the right hand side of (8). By the assumed H–Poincaré inequality (13), we obtain

$$-\frac{1}{2}\,\mathbb{E}_t\left[H_t(g_t-1)^2\right] \;\leq\; \frac{1}{2}\,\mathbb{E}_t\left[H_t^-(g_t-1)^2\right] \;\leq\; \frac{1}{2}A_t\cdot\mathcal{E}_t(g_t-1)\;.$$

Moreover,

$$\mathbb{E}_t\left[H_t(g_t-1)\right] \;\leq\; \left(\mathbb{E}_t[H_t^2]\right)^{1/2}\left(\mathbb{E}_t[(g_t-1)^2]\right)^{1/2} \;=\; \sigma_t(H)\,\varepsilon_t^{1/2}.$$

Substituting into (8) yields

$$\begin{aligned}\frac{d}{dt}\varepsilon_t &\leq -2\,(\lambda_t - A_t/2)\,\mathcal{E}_t(g_t-1) \,+\, 2\,\sigma_t(H)\,\varepsilon_t^{3/2}\\ &\leq -\frac{2\lambda_t - A_t}{C_t}\,\varepsilon_t \,+\, 2\,\sigma_t(H)\,\varepsilon_t^{3/2},\end{aligned}$$

by the global Poincaré inequality (12), provided $\lambda_t \geq A_t/2$. This proves (15).

On the other hand,

$$\begin{aligned}&\mathbb{E}_t\left[H_t\big(-(g_t-1)^2/2+(g_t-1)\varepsilon_t\big)\right]\\ &= \frac{1}{2}\,\mathbb{E}_t\left[H_t^-(g_t-1)^2\right] + \mathbb{E}_t\left[H_t^-(1-g_t)\right]\,\varepsilon_t\\ &\qquad + \mathbb{E}_t\left[H_t^+\big(-(g_t-1)^2/2+(g_t-1)\varepsilon_t\big)\right]\;.\end{aligned} \tag{17}$$

Estimating the three summands on the right hand side separately yields

$$\mathbb{E}_t\big[H_t^-(g_t-1)^2\big] \;\le\; A_t\cdot\mathcal{E}_t(g_t-1)$$

by the H-Poincaré inequality (13),

$$\begin{aligned}\mathbb{E}_t\big[H_t^-(1-g_t)\big] \;&\le\; \mathbb{E}_t[H_t^-]^{1/2}\,\mathbb{E}_t[H_t^-(g_t-1)^2]^{1/2}\\ &\le\; \mathbb{E}_t[H_t^-]^{1/2}A_t^{1/2}\mathcal{E}_t(g_t-1)^{1/2}\end{aligned}$$

by the Cauchy-Schwarz inequality and (13), and

$$\mathbb{E}_t\big[H_t^+\big(-(g_t-1)^2/2+(g_t-1)\varepsilon_t\big)\big] \;\le\; \frac{1}{2}\,\mathbb{E}_t[H_t^+]\,\varepsilon_t^2 \;=\; \frac{1}{2}\,\mathbb{E}_t[H_t^-]\,\varepsilon_t^2\ .$$

The last estimate follows since

$$\frac{1}{2}\,\varepsilon_t^2 \;\ge\; (g_t-1)\varepsilon_t-\frac{1}{2}(g_t-1)^2$$

and

$$\mathbb{E}_t[H_t^+]-\mathbb{E}_t[H_t^-] \;=\; \mathbb{E}_t[H_t] \;=\; 0\ .$$

By combining the estimates, (17) and (8), we obtain

$$\frac{d}{dt}\varepsilon_t \;\le\; -(2\lambda_t-A_t)\,\mathcal{E}_t(g_t-1)\;+\;2A_t^{1/2}\mathbb{E}_t[H_t^-]^{1/2}\mathcal{E}_t(g_t-1)^{1/2}\varepsilon_t\;+\;\mathbb{E}_t[H_t^-]\,\varepsilon_t^2\,.$$

This combined with the global Poincaré inequality (12) yields

$$\frac{d}{dt}\varepsilon_t \;\le\; -\frac{2\lambda_t-A_t}{C_t}\,\varepsilon_t\;+\;2\,\frac{A_t^{1/2}}{C_t^{1/2}}\,\mathbb{E}_t[H_t^-]^{1/2}\varepsilon_t^{3/2}\;+\;\mathbb{E}_t[H_t^-]\,\varepsilon_t^2\,,$$

and hence (16). □

As an immediate consequence of the theorem we obtain estimates on the average relative frequency λ_t of MCMC moves that is sufficient to guarantee stability of the corresponding nonlinear flow of probability measures.

Corollary 4. *Let $0\le\beta_0<\beta_1$, and assume that for all $t\in(\beta_0,\beta_1)$,*

$$\lambda_t \;>\; \frac{A_t}{2}\;+\;C_t\sigma_t(H)\,\varepsilon_{\beta_0}^{1/2} \tag{18}$$

or

$$\lambda_t \;>\; \frac{A_t}{2}\;+\;(A_tC_t\mathbb{E}_t[H_t^-])^{1/2}\,\varepsilon_{\beta_0}^{1/2}+\frac{1}{2}\,C_t\mathbb{E}_t[H_t^-]\,\varepsilon_{\beta_0}\ . \tag{19}$$

Then $t\mapsto\varepsilon_t$ is strictly decreasing on the interval $[\beta_0,\beta_1]$.

Proof. If (18) or (19) holds for $t\in(\beta_0,\beta_1)$, then by Theorem 3 and continuity, $t\mapsto\varepsilon_t$ is strictly decreasing near β_0 and near any $s\in(\beta_0,\beta_1)$ such that $\varepsilon_s\le\varepsilon_{\beta_0}$. Hence it is strictly decreasing on the whole interval $[\beta_0,\beta_1]$. □

Remark 5. (i) On the finite state spaces considered here, the constants C_t and A_t are finite if $\mathcal{L}_t$ is irreducible. However, in multimodal situations, the numerical values of these constants are often extremely large. Alternative estimates based on local Poincaré-type inequalities are in the next section.

(ii) Similarly to the above corollary, one obtains that the error decays exponentially with rate $\gamma > 0$, i.e. $t \mapsto e^{\gamma t}\,\varepsilon_t$ is decreasing on $[\beta_0, \beta_1]$, provided

$$\lambda_t \;>\; \frac{A_t + \gamma C_t}{2} + C_t \sigma_t(H)\, e^{-\gamma(t-\beta_0)/2}\, \varepsilon_{\beta_0}^{1/2} \qquad \forall\, t \in (\beta_0, \beta_1)\,, \tag{20}$$

or a similar condition replacing (19) holds.

(iii) One can often assume that the initial error ε_{β_0} is very small. In this case, λ_t slightly greater than $(A_t + \gamma C_t)/2$ is enough to ensure exponential decay with rate γ.

(iv) The case $H \equiv 0$ corresponds to classical MCMC. Here $A_t = 0$ for all t, so $\partial \varepsilon_t / \partial t \leq -2 \frac{\lambda_t}{C_t} \varepsilon_t$. This yields the classical exponential decay with rate 2γ of the mean square error in the presence of the global spectral gap $\lambda_t / C_t \geq \gamma$ of the generator $\lambda_t \cdot \mathcal{L}_t$. For $H \not\equiv 0$, additional MCMC moves are required to make up for the error growth due to importance sampling/resampling.

Roughly, the corollary says that is the initial error is sufficiently small, the stabilizing effects of the MCMC dynamics make up for the error growth due to importance sampling/resampling provided $\lambda_t \geq A_t/2$.

3.2 Comparison with parallel MCMC

Suppose that we want to simulate μ_β for a fixed $\beta > 0$. Parallel MCMC consists in simulating N independent time homogeneous Markov chains with generator $\mathcal{L}_\beta$. This algorithm is clearly a special case of the sequential MCMC procedure introduced above, where $\mu_t = \mu_\beta$ for all $t > 0$ and $H = 0$. If the chains are run with initial distribution μ_0, one has

$$\varepsilon_t \;\leq\; e^{-2t/C_\beta}\, \varepsilon_0 \;\leq\; e^{-2t/C_\beta} \cdot \left(e^{\beta\, \mathrm{osc}(H)} - 1 \right)$$

where we have used that

$$\varepsilon_0 \;=\; \sum_{x \in S} \left(\frac{\mu_0(x)}{\mu_\beta(x)} - 1 \right)^2 \mu_\beta(x) \;=\; \sum_{x \in S} \frac{\mu_0(x)}{\mu_\beta(x)}\, \mu_0(x) \;-\; 1 \;\leq\; e^{\beta\, \mathrm{osc}(H)} - 1\,.$$

Hence to ensure $\varepsilon_T < \bar{\varepsilon}$ for a given $\bar{\varepsilon} > 0$ and $T > 0$, a total running time

$$T \;\geq\; \frac{C_\beta}{2} \cdot \left(\beta\, \mathrm{osc}(H) + \log \frac{1}{\bar{\varepsilon}} \right)$$

is sufficient. If (**??**) holds, the number of MCMC steps required for a simulation is of the same order as T. Alternatively, we can apply the sequential MCMC method with varying distributions μ_t $(0 \leq t \leq \beta)$. Using the rough estimate $A_t \leq C_t \cdot \sup H_t^-$ and (19), we see that ε_t decreases in time if

$$\lambda_t \;\geq\; \frac{1}{2}\, C_t \sup H_t^- \, (1 + \varepsilon_0^{1/2})^2 \qquad \forall\, t \in (0, \beta).$$

Thus an expected total number of MCMC steps of order

$$\frac{1}{2}(1+\varepsilon_0^{1/2})^2 \int_0^\beta C_t \sup H_t^- \, dt$$

suffices to guarantee stability of the corresponding nonlinear semigroup.

More drastic improvements due to sequential MCMC appear when good global spectral gap estimates do not hold, as we shall now demonstrate.

4 Error control based on local estimates

Madras and Randall [8] and Jerrum, Son, Tetali and Vigoda [6] have shown how to derive estimates for spectral gaps and logarithmic Sobolev constants of the generator of a Markov chain from corresponding local estimates on the sets of a decomposition of the state space combined with estimates for the projected chain. This has been applied to tempering algorithms in [9], [1] and [**?**]. We now develop related decomposition techniques for sequential MCMC. However, in this case, we will assume *only* local estimates for the generators $\mathcal{L}_t$, and no mixing properties for the projections – whence there will be an unavoidable error growth due to importance sampling/resampling between the components. The results and examples below indicate that nevertheless sequential MCMC methods might potentially be at least equally efficient as tempering algorithms in many applications. Since mixing properties for the projections do not have to be taken into account, the analysis of the decomposition simplifies considerably.

Let $0 \le \beta_0 < \beta_1 \le \infty$. We assume that for every $t \in (\beta_0, \beta_1)$, there exists a decomposition

$$S = \bigcup_{i \in I} S_t^i$$

into finitely many disjoint sets with $\mu_t(S_t^i) > 0$, as well as non–negative definite quadratic forms $\mathcal{E}_t^i$ $(i \in I)$ on functions on S such that

$$\sum_i \mu_t(S_t^i)\, \mathcal{E}_t^i(f) \;\le\; K \cdot \mathcal{E}_t(f) \qquad \forall\, t \in (\beta_0, \beta_1),\; f : S \to \mathbb{R} \tag{21}$$

for some fixed finite constant K. For example, one might choose $\mathcal{E}_t^i$ as the Dirichlet form of the Markov chain corresponding to $\mathcal{L}_t$ restricted to S_t^i, i.e.,

$$\mathcal{E}_t^i(f) \;=\; \frac{1}{2} \sum_{x,y \in S_t^i} (f(y) - f(x))^2\, \mathcal{L}_t(x,y)\, \mu_t(x \,|\, S_t^i)\,. \tag{22}$$

In this case, (21) holds with $K = 1$.

Let us denote by $\mathbb{E}_t^i$ and Var_t^i, respectively, the expectation and variance w.r.t. the conditional measure

$$\mu_t^i(A) := \mu_t(A | S_t^i),$$

and by $\pi : S \to I$ the natural projection. In particular,

$$\mathbb{E}_t[f|\pi] = \sum_{i \in S} \mathbb{E}_t^i[f] \cdot \chi_{S_t^i},$$

for any function $f : S \to \mathbb{R}$. We set

$$\tilde{H}_t := H_t - \mathbb{E}_t[H_t|\pi].$$

Assume that the following **local Poincaré-type inequalities** hold for all $t \in (\beta_0, \beta_1)$ and $i \in I$ with constants $A_t^i, B_t^i \in (0, \infty)$:

$$\mathbb{E}_t^i[-\tilde{H}_t\, f^2] \;\leq\; A_t^i \cdot \mathcal{E}_t^i(f) \qquad \forall f : S \to \mathbb{R} \,:\, \mathbb{E}_t[f|\pi] = 0 \;, \tag{23}$$

$$\left|\mathbb{E}_t^i[\tilde{H}_t\, f]\right|^2 \;\leq\; B_t^i \cdot \mathcal{E}_t^i(f) \qquad \forall f : S \to \mathbb{R} \,:\, \mathbb{E}_t[f|\pi] = 0 \;. \tag{24}$$

Remark 6. (i) Note that to verify (23) it is enough to estimate $\mathbb{E}_t^i[\tilde{H}_t^- f^2]$, while for (24) one has to take into account the positive part of $\tilde{H}_t$ as well. In particular, (23) can not be used to derive an estimate of type (24). However, if (23) holds with $-\tilde{H}_t$ replaced by $|\tilde{H}_t|$, then (24) holds with $B_t^i = \mathbb{E}_t^i[|\tilde{H}_t|] \cdot A_t^i$.

(ii) If local Poincaré inequalities of the type

$$\mathrm{Var}_t^i(f) \;\leq\; C_t^i \cdot \mathcal{E}_t^i(f) \qquad \forall\, f : S \to \mathbb{R},\; i \in I, \tag{25}$$

hold, then (23) and (24) hold with $A_t^i = C_t^i \cdot \max_{S_i} \tilde{H}_t^-$ and $B_t^i = C_t^i \cdot \mathrm{Var}_t^i(H)$.

Combining (21) and (23), (24) respectively yields

$$\mathbb{E}_t[-\tilde{H}_t \tilde{f}_t^2] \;=\; \sum_{i \in I} \mu_t(S_t^i)\, \mathbb{E}_t^i[-\tilde{H}_t \tilde{f}_t^2] \;\leq\; \hat{A}_t \cdot \mathcal{E}_t(f) \qquad \forall\, f : S \to \mathbb{R}\,, \tag{26}$$

and

$$\sum_{i \in I} \mu_t(S_t^i) \left|\mathbb{E}_t^i[\tilde{H}_t \tilde{f}_t]\right|^2 \;\leq\; \hat{B}_t \cdot \mathcal{E}_t(f) \qquad \forall\, f : S \to \mathbb{R}\,, \tag{27}$$

where

$$\hat{A}_t := K \cdot \max_i A_t^i \qquad \text{and} \qquad \hat{B}_t := K \cdot \max_i B_t^i.$$

The following error estimate is our key result.

Theorem 7. *If $\lambda_t > \hat{A}_t/2$ for all $t \in (\beta_0, \beta_1)$ then*

$$\frac{d}{dt} \log \varepsilon_t \;\leq\; \frac{\hat{B}_t}{\lambda_t - \hat{A}_t/2} \cdot (1 + \varepsilon_t) \;+\; (1 + \sqrt{\varepsilon_t})^2 \cdot \max_{i \in I} h_t^-(i) \tag{28}$$

where

$$h_t(i) := \mathbb{E}_t^i[H_t] = -\left.\frac{\partial}{\partial s} \log \mu_s(S_t^i)\right|_{s=t} \qquad (i \in I). \tag{29}$$

Proof. Similarly to Theorem 3, we have to control the right hand side of (8), but now by using only local Poincaré type inequalities. Let

$$f_t := g_t - 1 \qquad \text{and} \qquad \tilde{f}_t := f_t - \mathbb{E}_t[f_t|\pi].$$

Then

$$
\begin{aligned}
&\mathbb{E}_t\left[H_t\big(-(g_t-1)^2/2+(g_t-1)\varepsilon_t\big)\right] \\
&\quad= \mathbb{E}_t\left[\tilde{H}_t\big(-(g_t-1)^2/2+(g_t-1)\varepsilon_t\big)\right] \\
&\qquad + \sum_{i\in I}\mu_t(S_t^i)\,\mathbb{E}_t^i[H_t]\,\mathbb{E}_t^i\left[-(g_t-1)^2/2+(g_t-1)\varepsilon_t\right] \\
&\quad= -\frac{1}{2}\,\mathbb{E}_t[\tilde{H}_t\tilde{f}_t^2]-\mathbb{E}_t\big[\tilde{H}_t\tilde{f}_t\,\mathbb{E}_t[f_t|\pi]\big]+\mathbb{E}_t[\tilde{H}_t\tilde{f}_t\,\varepsilon_t] \\
&\qquad + \sum_{i\in I}\mu_t(S_t^i)\,\mathbb{E}_t^i[H]\,\mathbb{E}_t^i\left[-f_t^2/2+f_t\varepsilon_t\right] \qquad (30)\\
&\quad= -\frac{1}{2}\,\mathbb{E}_t[\tilde{H}_t\tilde{f}_t^2]\;+\sum_{i\in I}\mu_t(S_t^i)\,\mathbb{E}_t^i[\tilde{H}_t\tilde{f}_t]\,\big(\varepsilon_t-\mathbb{E}_t^i[f_t]\big) \\
&\qquad + \sum_{i\in I}\mu_t(S_t^i)\,h_t(i)\,\mathbb{E}_t^i[-f_t^2/2+f_t\varepsilon_t].
\end{aligned}
$$

Here we have used the definitions of H_t, $\tilde{H}_t$ and h_t, and the fact that $\mathbb{E}_t[\tilde{H}_t|\pi]=0$. We now estimate the three summands on the right hand side separately. By the local H-Poincaré inequality (26),

$$
-\frac{1}{2}\,\mathbb{E}_t[\tilde{H}_t\tilde{f}_t^2]\;\leq\frac{1}{2}\,\hat{A}_t\cdot\mathcal{E}_t(f_t).
$$

By (27), and since

$$
\sum_i\mu_t(S_t^i)\,\mathbb{E}_t^i[f_t]=\mathbb{E}_t[f_t]=0,
$$

we have

$$
\begin{aligned}
&\sum_{i\in I}\mu_t(S_t^i)\cdot\mathbb{E}_t^i[\tilde{H}_t\tilde{f}_t]\cdot\big(\varepsilon_t-\mathbb{E}_t^i[f_t]\big) \\
&\quad\leq \left(\sum_{i\in I}\mu_t(S_t^i)\,\mathbb{E}_t^i[\tilde{H}_t\tilde{f}_t]^2\right)^{1/2}\left(\sum_{i\in I}\mu_t(S_t^i)\,(\varepsilon_t-\mathbb{E}_t^i[f_t])^2\right)^{1/2} \\
&\quad\leq \hat{B}_t^{1/2}\mathcal{E}_t(f_t)^{1/2}\cdot\left(\varepsilon_t^2+\sum\mu_t(S_t^i)\mathbb{E}_t^i[f_t^2]\right)^{1/2} \\
&\quad= \left(\hat{B}_t\,\mathcal{E}_t(f_t)\cdot\varepsilon_t\cdot(1+\varepsilon_t)\right)^{1/2}.
\end{aligned}
$$

Moreover, since

$$
-f_t^2/2+f_t\varepsilon_t\;\leq\;\varepsilon_t^2/2,
$$

we obtain

$$
\begin{aligned}
&\sum_{i\in I}\mu_t(S_t^i)\,h_t(i)\,\mathbb{E}_t^i[-f_t^2/2+f_t\varepsilon_t] \\
&\quad\leq \sum_{i\in I}\mu_t(S_t^i)\,h_t^+(i)\cdot\frac{1}{2}\,\varepsilon_t^2\;+\sum_{i\in I}\mu_t(S_t^i)\,h_t^-(i)\,\mathbb{E}_t^i[f_t^2/2-f_t\varepsilon_t] \\
&\quad\leq \left(\frac{1}{2}\,\varepsilon_t^2\;+\;\frac{1}{2}\,\varepsilon_t\;+\;\varepsilon_t^{3/2}\right)\cdot\max h_t^-\;=\;\varepsilon_t\cdot(1+\sqrt{\varepsilon_t})^2\cdot\max h_t^-\,.
\end{aligned}
$$

Here we have used that

$$\sum \mu_t(S_t^i) h_t^+(i) = \sum \mu_t(S_t^i) h_t^-(i) \leq \max h_t^-$$

and

$$\sum \mu_t(S_t^i)\, \mathbb{E}_t^i[-f_t] \leq \left(\sum \mu_t(S_t^i)\, \mathbb{E}_t^i[f_t^2]\right)^{1/2} = \varepsilon_t^{1/2}.$$

Combining the estimates yields, by (8) and (30):

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\varepsilon_t \;&\leq\; -\lambda_t \cdot \mathcal{E}_t(f_t) \;+\; \mathbb{E}_t\big[H_t(-f_t^2/2 + f_t\varepsilon_t)\big] \\
&\leq\; -\left(\lambda_t - \frac{\hat{A}_t}{2}\right) \cdot \mathcal{E}_t(f_t) \;+\; \left(\hat{B}_t \mathcal{E}_t(f_t)\varepsilon_t(1+\varepsilon_t)\right)^{1/2} \;+\; \frac{1}{2}\,\varepsilon_t(1+\sqrt{\varepsilon_t})^2 \max h_t^- \\
&\leq\; \frac{\hat{B}_t}{2\lambda_t - \hat{A}_t}\,\varepsilon_t\,(1+\varepsilon_t) \;+\; \frac{1}{2}\,\varepsilon_t\;(1+\sqrt{\varepsilon_t})^2 \max h_t^- \;,
\end{aligned}$$

provided $\lambda_t > \hat{A}_t/2$. This proves (28).

Moreover, for any subset $A \subseteq S$,

$$\begin{aligned}
\frac{d}{dt}\log \mu_t(A) &= \frac{1}{\mu_t(A)} \sum_{x\in A} \partial_t \mu_t(x) = \frac{1}{\mu_t(A)} \sum_{x\in A} \big(\partial_t \log \mu_t(x)\big)\mu_t(x) \\
&= \frac{1}{\mu_t(A)} \sum_{x\in A} H_t(x)\mu_t(x) = \mathbb{E}_t[H_t|A],
\end{aligned}$$

which proves (29). □

To understand the consequences of (28), let us first consider the asymptotics as λ_t tends to infinity. In this case, (28) reduces to

$$\frac{d}{dt}\log \varepsilon_t \;\leq\; (1+\sqrt{\varepsilon_t})^2 \cdot \max h_t^- \;.$$

In order to ensure that for $t > \beta_0$ the error ε_t remains below a given threshold $\delta > 0$, note that as long as $\varepsilon_t \leq \delta$, we have

$$\frac{d}{dt}\log \varepsilon_t \leq \left(1+\sqrt{\delta}\right)^2 \max h_t^-.$$

Thus

$$\min(\varepsilon_t, \delta) \leq \varepsilon_{\beta_0} G_t^{(1+\sqrt{\delta})^2} \qquad \forall t \in [\beta_0, \beta_1], \tag{31}$$

where

$$G_t := \exp\left(\int_{\beta_0}^t \max h_r^- \, dr\right) = \exp\left(\int_{\beta_0}^t \max_i \frac{\partial}{\partial s}\log \mu_s(S_r^i)\bigg|_{s=r} dr\right).$$

Remark 8. The term $G_t^{(1+\sqrt{\delta})^2}$ in (31) accounts for the maximum error growth due to importance sampling between the components. If $S_t^i = S^i$ is independent of t for every

i, and there is an $i_0 \in I$ such that $\partial_s \log \mu_s(S^i)$ is maximized by S^{i_0} for all $s \in (\beta_0, \beta_1)$, then

$$G_t \;=\; \exp\left(\int_{\beta_0}^t \max_i \frac{d}{ds} \log \mu_s(S^i)\, ds\right) \;=\; \frac{\mu_t(S^{i_0})}{\mu_{\beta_0}(S^{i_0})} \qquad \forall\; t \in [\beta_0, \beta_1],$$

i.e., G_t is the growth rate of this strongest growing component. In general, things are more complicated, but a similar interpretation is at least possible on appropriate subintervals of $[\beta_0, \beta_1]$.

Now we return to the case when λ_t is finite. The next corollary tells us how many MCMC moves are sufficient to obtain an estimate on the growth of ε_t that is not much worse than (31).

Corollary 9. *Let $\beta \in (\beta_0, \beta_1]$ and $\delta > 0$, and assume that*

$$\lambda_t \geq \frac{\hat{A}_t}{2} + \alpha_t \cdot \hat{B}_t \qquad \forall t \in (\beta_0, \beta) \tag{32}$$

for some function $\alpha : (\beta_0, \beta) \to (0, \infty)$. Then

$$\min(\varepsilon_\beta, \delta) \leq \varepsilon_{\beta_0} G_\beta^{(1+\sqrt{\delta})^2} \exp \int_{\beta_0}^{\beta} \frac{1+\delta}{\alpha_s}\, ds \qquad \forall t \in [\beta_0, \beta]. \tag{33}$$

In particular, if

$$\lambda_t \geq \frac{\hat{A}_t}{2} + (\beta - \beta_0)\hat{B}_t \qquad \forall t \in (\beta_0, \beta), \tag{34}$$

then

$$\min(\varepsilon_\beta, \delta) \leq \varepsilon_{\beta_0} G_\beta^{(1+\sqrt{\delta})^2} e^{1+\delta}. \tag{35}$$

Proof. Assume that (32) holds, and let

$$u_t := \varepsilon_t / G_t^{(1+\sqrt{\delta})^2}.$$

Then by the definition of G_t, Theorem 7, and (32),

$$\begin{aligned} \frac{d}{dt} \log u_t &= \frac{d}{dt} \log \varepsilon_t - (1+\sqrt{\delta})^2 \max h_t^- \\ &\leq \frac{\hat{B}_t}{\lambda_t - \hat{A}_t/2}(1+\delta) \leq \frac{1+\delta}{\alpha_t} \end{aligned}$$

for all $t \in (\beta_0, \beta)$ such that $\varepsilon_t \leq \delta$. Hence

$$\varepsilon_t = u_t\, G_t^{(1+\sqrt{\delta})^2} \leq \varepsilon_{\beta_0} G_t^{(1+\sqrt{\delta})^2} \exp \int_{\beta_0}^t \frac{1+\delta}{\alpha_s}\, ds$$

holds for $t \in [\beta_0, \beta]$ provided the right hand side is smaller than δ. This proves (33). The second assertion is a straightforward consequence. □

Remark 10. The main difference to Corollary 4 is that under local conditions it can not be guaranteed that the error remains bounded. Instead, ε_t can grow with a rate dominated by $G_t^{(1+\sqrt{\delta})^2}$. As already pointed out, this is due to importance sampling between the components.

5 Examples

5.1 Exponential model with k valleys in the energy landscape

This is an extended version of a model considered in [7], [9] as a test case for some multi-level MCMC methods. We fix $k \in \mathbb{N}$, and $r_1, r_2, \ldots, r_k \in \mathbb{N}$. Let $S^0 := \{0\}$ and

$$S^i \;:=\; \{(i,j) : j = 1, 2, \ldots, r_i\}, \qquad 1 \le i \le k.$$

We consider the graph with vertex set

$$S \;=\; \bigcup_{i=0}^{k} S^i$$

and edges $(0,(i,1))$, $1 \le i \le k$, and $((i,j),(i,j+1))$, $1 \le i \le k$, $1 \le j \le r_i - 1$. Suppose that

$$H(x) \;=\; -d(x,0), \qquad x \in S,$$

where $d(x,0)$ stands for the graph distance of x from 0, i.e., $H(0) = 0$ and $H((i,j)) = -j$. We assume that μ_t is given by (**??**), where μ is an arbitrary probability distribution on S such that $\mu(x) > 0$ for all $x \in S$ and μ is log-concave on each of the valleys S^i of the energy landscape, i.e.,

$$\frac{1}{2}\big(\log \mu((i,j+1)) + \log \mu((i,j-1))\big) \le \log \mu((i,j))$$

for all $1 \le i \le k$ and $1 \le j \le r_i$. We consider the setup for sequential MCMC as described above where $\mathcal{L}_t$ is the generator of the Metropolis dynamics w.r.t. μ_t based on the nearest neighbor random walk on S. Of course, there are more efficient ways to carry out Monte Carlo integrations in this special situation. The point, however, is that sequential MCMC methods can be applied even though the underlying structure of the energy landscape is unknown. Let $R = \max_{1 \le i \le k} r_i$. An application of Corollary 9 with $\beta_0 = 0$ and $S_t^i = S^i$ for all $t \ge 0$ yields the following result :

Theorem 11. *If*

$$\lambda_t \ge R^3 + \frac{\beta}{2} R^4 \qquad \forall t \in (0,\beta),$$

then

$$\min(\varepsilon_\beta, \delta) \;\le\; e^{1+\delta} \cdot \varepsilon_0 G_\beta^{(1+\sqrt{\delta})^2} \cdot \varepsilon_0 \qquad \forall\ \delta \in (0,1). \tag{36}$$

Moreover, if the conditional distribution $\mu(\cdot|S^{i_0})$ lies deeper in one of the valleys than in the others in the sense that

$$\mu\big(\{(i,j) : j \ge h\}\big|S^{i_0}\big) \;\ge\; \mu\big(\{(i,j) : j \ge h\}\big|S^{i}\big)\ , \tag{37}$$

then

$$G_\beta \;=\; \frac{\mu_\beta(S^{i_0})}{\mu(S^{i_0})}\ ,$$

and thus

$$\min(\varepsilon_\beta, \delta) \;\le\; e^{1+\delta} \cdot \frac{\varepsilon_0}{\mu(S^{i_0})^{(1+\sqrt{\delta})^2}} \qquad \forall\ 0 < \delta < 1. \tag{38}$$

Proof. The log-concavity of μ easily implies that μ_t as well is log-concave on S^i for all $t \geq 0$ and $1 \leq i \leq k$. In particular, the restriction of μ_t to S^i has a unique local maximum for every i. By the method of canonical paths it is then not difficult to prove that the spectral gap of the Metropolis dynamics w.r.t. $\mu_t(\cdot|S^i)$ based on the standard random walk is bounded from below by $1/2r_i^2$ for all $t \geq 0$ and $1 \leq i \leq k$, cf. e.g. Proposition 6.3 in [5]. Now we are in the setting of Remark 6 (ii), according to which (23) and (24) hold with $\mathcal{E}_t^i$ as in (22),

$$A_t^i = 2r_i^3 , \qquad \text{and} \qquad B_t^i = \frac{1}{2} r_i^4 \ .$$

Estimate (36) now follows by a straightforward application of Corollary 9.

To prove the second part of the assertion, we show that (37) places us in the setting of Remark 8. In fact, for $t > 0$,

$$\frac{d}{dt} \log \mu_t(S^i) = \mathbb{E}_t[H] - \mathbb{E}_t^i[H] \qquad \text{for all } i$$

and

$$-\mathbb{E}_t^i[H] \ = \ -\frac{\mu\left(He^{-tH} \middle| \, S_i\right)}{\mu\left(e^{-tH} \middle| \, S_i\right)} \ = \ \frac{\sum_j j e^{tj} \mu((i,j))}{\sum_j e^{tj} \mu((i,j))} \ .$$

If (37) holds, then for any $t > 0$, the right hand side is maximized when $i = i_0$. Hence by Remark 8,

$$G_t \ = \ \frac{\mu_t(S_{i_0})}{\mu(S_{i_0})} \qquad \text{for all } t \geq 0.$$

□

Remark 12. (i) The last estimate indicates that to obtain good bounds it is crucial that the mass allocated by the initial distribution on the component S^{i_0} with strongest importance growth is not too small (although it can be rather small if the initial distribution ν_0 is a good approximation of μ_0).

(ii) Let $K_\beta = \int_0^\beta \lambda_t \, dt$. Note that K_β is a measure for the total number of MCMC steps that a corresponding sequential MCMC algorithm will perform on average. The theorem implies that choosing λ_t constant on $[0, \beta]$ with K_β of order $O(\beta^2)$ is sufficient to guarantee that the nonlinear flow of measures has good stability properties on $[0, \beta]$, and can thus be used to efficiently approximate μ_β. In contrast to this situation, the flow of measures corresponding to the standard simulated annealing algorithm has good stability properties only if K_β grows exponentially in β.

5.2 The mean field Ising model

As a very simple example for a model with a phase transition, we now consider the mean field Ising (Curie–Weiss) model, i.e. μ_β is of type (**??**) where $\mu_0 = \mu$ is the uniform distribution on the hypercube

$$S \ = \ \{-1, +1\}^N ,$$

and

$$H(\sigma) \ = \ -\frac{1}{2N} \sum_{i,j=1}^{N} \sigma_i \sigma_j \tag{39}$$

for some $N \in \mathbb{N}$. Let $\mathcal{L}_\beta$ be the generator of the (time–continuous) Metropolis chain w.r.t. μ_β based on the nearest neighbor random walk on S as proposal matrix. It is well known that this chain is rapidly mixing (i.e. the spectral gap decays polynomially in N) for $\beta < 1$, but torpid mixing holds (i.e. the spectral gap decays exponentially fast in N) for $\beta > 1$. Thus in the multi-phase regime $\beta > 1$, the classical Metropolis algorithm converges to equilibrium extremely slowly for large N.

Now assume for simplicity that N is odd, and decompose S into the two components

$$S^+ \;:=\; \left\{ \sigma \in S \;\middle|\; \sum_{i=1}^N \sigma_i > 0 \right\} \qquad \text{and}$$
$$S^- \;:=\; \left\{ \sigma \in S \;\middle|\; \sum_{i=1}^N \sigma_i < 0 \right\}.$$

Improving on previous results (e.g. of Madras and Zheng [9]), Schweizer [**?**] showed recently that the spectral gaps of the restricted Metropolis chains on both S^+ and S^- are bounded from below by $\frac{1}{9}N^{-2}$ for *every* $t \geq 0$. Applying the results above to the error growth for the non-linear semigroup corresponding to sequential MCMC in this situation, we obtain :

Theorem 13. *For every $\beta > 0$ and $N \in \mathbb{N}$,*

$$\sup_{0 \leq t \leq \beta} \varepsilon_t \;\leq\; e^2 \cdot \varepsilon_0$$

holds whenever $\varepsilon_0 \leq 1$ and

$$\lambda_t \;\geq\; \frac{9}{4} N^3 + \frac{9}{8} \beta N^4 \qquad \forall\, t \in (0, \beta). \tag{40}$$

Proof. Since $-N/2 \leq H(\sigma) \leq 0$ for all σ, we have $\operatorname{osc}(H) \leq N/2$ and

$$\mathrm{Var}_t(H|S^+) \;=\; \mathrm{Var}_t(H|S^-) \;\leq\; \left(\frac{1}{2} \operatorname{osc}(H) \right)^2 \;\leq\; N^2/8$$

for every $t \geq 0$. By Schweizer's result [**?**], a local Poincaré inequality of type (25) holds on S^+ and S^- with $C_t^+ = C_t^- = 9N^2$. Hence by Remark 6 (ii), (23) and (24) hold with

$$A_t^\pm \;=\; \frac{9}{2} N^3 \qquad \text{and} \qquad B_t^\pm \;=\; \frac{9}{8} N^4 .$$

The assertion now follows from Corollary 9, since

$$\mathbb{E}_t^+[H] \;=\; \mathbb{E}_t^-[H] \;=\; \mathbb{E}_t[H] .$$

□

Remark 14. (i) The result is based on a rough estimate of $\hat{A}_t$ and $\hat{B}_t$ in terms of the local spectral gap. We expect that a more precise estimate of these constants would yield a smaller power of N in (40). Furthermore, for $\beta \leq 1$, the result can be improved by applying global instead of local spectral gap estimates. However, our main interest is the phase transition regime.

(ii) Related results for the mean field Ising model have been obtained for mixing times of Markov chains for umbrella sampling in [7], and for simulated and parallel tempering in [9], [1], [**?**]. Schweizer [**?**] obtains an upper bound on the order in N and β of the L^2 mixing time (inverse spectral gap) for simulated tempering that is close to the one in (40). In contrast, the best known order for parallel tempering is much worse. In general, it seems that the analysis of sequential MCMC is partially simpler than the one for parallel tempering, where one has to take into account that a particle can only move in temperature if another particle moves in the opposite direction. In fact, for this reason we would expect that sequential MCMC methods can have substantial advantages compared to parallel tempering.

(iii) The theorem can be extended to a mean field Ising model with magnetic field. In this case, however, one has to take into account an additional (but well controlled) error growth due to importance sampling/resampling between the components. Moreover, the decomposition into the two components will now depend on t. Without magnetic field this is not the case because of the built-in symmetry.

References

[1] N. Bhatnagar and D. Randall, *Torpid mixing of simulated tempering on the Potts model*, SODA '04: Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms (Philadelphia, PA, USA), Society for Industrial and Applied Mathematics, 2004, pp. 478–487.

[2] P. Diaconis and L. Saloff-Coste, *Comparison theorems for reversible Markov chains*, Ann. Appl. Probab. **3** (1993), no. 3, 696–730. MR MR1233621 (94i:60074)

[3] ———, *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Probab. **6** (1996), no. 3, 695–750. MR MR1410112 (97k:60176)

[4] ———, *Nash inequalities for finite Markov chains*, J. Theoret. Probab. **9** (1996), no. 2, 459–510. MR MR1385408 (97d:60114)

[5] ———, *What do we know about the Metropolis algorithm?*, J. Comput. System Sci. **57** (1998), no. 1, 20–36, 27th Annual ACM Symposium on the Theory of Computing (STOC'95) (Las Vegas, NV). MR MR1649805 (2000b:68094)

[6] M. Jerrum, J.-B. Son, P. Tetali, and E. Vigoda, *Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains*, Ann. Appl. Probab. **14** (2004), no. 4, 1741–1765. MR MR2099650 (2005i:60139)

[7] N. Madras and M. Piccioni, *Importance sampling for families of distributions*, Ann. Appl. Probab. **9** (1999), no. 4, 1202–1225. MR MR1728560 (2001e:60139)

[8] N. Madras and D. Randall, *Markov chain decomposition for convergence rate analysis*, Ann. Appl. Probab. **12** (2002), no. 2, 581–606. MR MR1910641 (2003d:60135)

[9] N. Madras and Z. Zheng, *On the swapping algorithm*, Random Structures Algorithms **22** (2003), no. 1, 66–97. MR MR1943860 (2004c:82117)

[10] L. Saloff-Coste, *Lectures on finite Markov chains*, Lectures on probability theory and statistics (Saint-Flour, 1996), Lecture Notes in Math., vol. 1665, Springer, Berlin, 1997, pp. 301–413. MR MR1490046 (99b:60119)

Andreas Eberle
Institut für Angewandte Mathematik
Universität Bonn
Endenicher Str. 60, 53115 Bonn, Germany.

Carlo Marinelli
Facoltà di Economia
Università di Bolzano
39100 Bolzano, Italy.