

CONVERGENCE OF SEQUENTIAL MARKOV CHAIN MONTE CARLO METHODS: I. NONLINEAR FLOW OF PROBABILITY MEASURES

ANDREAS EBERLE AND CARLO MARINELLI

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.

Spectral gap estimates, or, equivalently, Poincaré inequalities, as well as other related functional inequalities provide powerful tools for the study of convergence to equilibrium of reversible time-homogeneous Markov processes (see e.g. [9], [10], [11]). In particular, they have been successfully applied to analyze convergence properties of Markov Chain Monte Carlo (MCMC) methods based on reversible Markov chains (see e.g. [12]). The idea of MCMC methods is to produce approximate samples from a probability distribution μ by simulating for a sufficiently long time an ergodic Markov chain having μ as invariant measure. MCMC methods have become the standard to carry out Monte Carlo integrations with respect to complex probability distributions in many fields of applications, including in particular Bayesian statistics, statistical physics, and computational chemistry. We just refer the interested reader to [15] and [19] and references

Date: December 3, 2006.

2000 Mathematics Subject Classification. 65C05, 60J25, 60B10, 47H20, 47D08.

Key words and phrases. Markov Chain Monte Carlo, sequential Monte Carlo, importance sampling, spectral gap, Dirichlet forms, functional inequalities.

This work is supported by the Sonderforschungsbereich 611, Bonn. The second author also gratefully acknowledges the hospitality and financial support of the Institute of Mathematics, Polish Academy of Sciences, Warsaw, of the Institute des Hautes Études Scientifiques, Bures-sur-Yvette, of the Max-Planck-Institut für Mathematik, Leipzig, through an IPDE fellowship.

therein as an example of work in this area, as the literature is by now enormous. Since the Markov chain is usually started with an initial distribution that is very different from μ , strong convergence properties, such as exponential convergence to equilibrium with a sufficiently large rate, are required to ensure that the corresponding MCMC method produces sufficiently good approximate samples from μ . However, these strong convergence properties often do not hold in multimodal, and in particular high-dimensional problems, as they arise in many applications. For example, in statistical mechanics models with phase transitions, the rate of convergence often decays exponentially in the system size within the multi-phase regime.

In this and a follow-up article we initiate a study of convergence properties by functional inequalities for a class of algorithms for Monte Carlo integral estimation that are a combination of sequential Monte Carlo and MCMC methods. Instead of producing constantly improved samples of a fixed distribution μ , these *sequential MCMC methods* try to keep track as precisely as possible of an evolving sequence $(\mu_t)_{0 \leq t \leq \beta}$ of probability distributions. Here μ_0 is an initial distribution that is easy to simulate, and μ_β is the target distribution that one would like to simulate. Importance sampling and resampling steps are included to constantly adjust for the change of the underlying measure. Whereas for MCMC methods exponential asymptotic stability is usually required to obtain improved samples, the sequential MCMC method starts with a good estimate of μ_0 , and one only has to control the growth of the “size” of the error. As a consequence, the method sometimes works surprisingly well in multimodal situations where traditional MCMC methods fail, cf. also the examples below. The price one has to pay is that samples from μ_β cannot be produced individually. Instead, the corresponding algorithm produces directly a Monte Carlo estimator for μ_β given by the empirical distribution of a system of interacting particles at the final time. To ensure good approximation properties, a large number N of particles is required.

Variants of such sequential MCMC methods have recently been proposed at several places in the statistics literature, see in particular [6] and references therein, as well as [3]. However, precise and general mathematical methods for the convergence and stability analysis, in the spirit of those developed for traditional MCMC methods by Diaconis, Saloff-Coste, Jerrum, Sinclair, and many others, seem still to be missing – although very important first steps can be found in the work of Del Moral and coauthors, cf. e.g. [5], [7] and [8]. The classical approach via Dobrushin contraction coefficients is usually limited to very regular situations. Moreover, it rarely yields precise statements

on the convergence properties, and it can not be combined easily with decomposition techniques.

Our aim is to make variants of the powerful techniques of the spectral gap/Dirichlet form approach to convergence rates of time-homogeneous Markov chains (e.g. canonical paths, comparison and decomposition results) available in the different context of sequential MCMC methods. Mathematically, this means at first to study a class of nonlinear evolutions of probability measures by functional inequalities. Such a study has been initiated in a related context by Stannat [22]. In this work, we restrict ourselves to the simplest and most natural variant of sequential MCMC, where importance sampling/resampling is *only* used to adjust constantly for the change of the underlying distribution, and MCMC steps at time t are always carried out such that detailed balance holds w.r.t. the measure μ_t (and not w.r.t. μ_0 !). This seems crucial for establishing good stability properties. Note that the type of sequential Monte Carlo samplers studied here is different from those analyzed by Del Moral and Doucet in [5]. An algorithmic realization has been applied to simulations in Bayesian mixture models by Del Moral, Doucet and Jasra in [6], who observed substantial benefits compared to other methods.

We have divided our work on sequential MCMC methods into two publications: in this first article we study the stability properties of nonlinear flows of probability measures describing the limit as the number N of particles goes to infinity. In the follow-up work [13] we will apply the results to control the asymptotic variances of the Monte Carlo estimators as N tends to infinity. The functional inequality approach enables us to prove stability properties not only under global but also under local conditions, i.e. assuming only that good estimates hold on each set of a decomposition of the state space. As a consequence, we obtain a procedure for analyzing the asymptotic behavior of sequential MCMC methods applied to multimodal distributions. For example, in the spirit of previous results for tempering algorithms by Madras and Zheng [18] and others, we can prove good (polynomial in the system size and the inverse temperature) stability properties in the case of the mean field Ising model, cf. Section 2.5 below. We also demonstrate in a simple exponential model with several modes that the flow of probability measures corresponding to sequential MCMC methods has better stability properties than the one corresponding to the classical simulated annealing algorithm, cf. 2.4. Sequential MCMC methods might hence also provide an efficient alternative to simulated annealing.

1. SETUP

1.1. Sequential estimation of probability measures. Let S denote a finite state space, and μ a probability measure on S with full support, i.e. $\mu(x) > 0$ for all $x \in S$. The finiteness of the state space is only assumed to keep the presentation as simple and non-technical as possible. Most results of this paper extend to continuous state spaces under standard regularity assumptions. By $\mathcal{M}_1(S)$ we denote the space of probability measures on S . As usual,

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

denotes the expectation of a function $f : S \rightarrow \mathbb{R}$ w.r.t. a measure $\nu \in \mathcal{M}_1(S)$. We consider methods for Monte Carlo integration with respect to the probability distributions of an exponential family

$$\mu_t(x) = \frac{1}{Z_t} e^{-tH(x)} \mu(x), \quad 0 \leq t < \infty, \quad (1)$$

where $H : S \rightarrow [0, \infty)$ is a given function, and $Z_t := \sum_{x \in S} e^{-tH(x)} \mu(x)$ is a normalization constant. Below, t will play the rôle of a time parameter for a particle system approximation.

Note that for a fixed $\beta > 0$, *any* given probability measure ν on S that is mutually absolutely continuous with respect to μ can be written in the form (1) with $t = \beta$ by setting $H(x) = \frac{1}{\beta} \log \frac{\mu(x)}{\nu(x)}$. One should then think of the family $(\mu_t)_{0 \leq t \leq \beta}$ of probability measures as a particular way to interpolate between the target distribution μ_β that we would like to simulate, and the reference distribution $\mu_0 = \mu$ that can be simulated more easily. Although we restrict our attention here to this simple way of interpolating between two measures, other interpolations can be treated by similar methods. In fact, an arbitrary family $(\mu_t)_{0 \leq t \leq \beta}$ of mutually absolutely continuous probability measures on S with smooth dependence on t can be written in the form

$$\mu_t(x) = \frac{1}{Z_t} e^{-\int_0^t U_s(x) ds} \mu(x), \quad 0 \leq t \leq \beta, \quad (2)$$

where Z_t is a normalization constant, and $(s, x) \mapsto U_s(x)$ is a continuous non-negative function on $[0, \beta] \times S$. Our results below extend to this more general case, cf. Section 2.6 below.

The main advantage of the interpolation (1) is that the singularity of μ_t w.r.t. μ is resolved uniformly over time. In particular,

$$\left| \log \frac{\mu_s(x)}{\mu_t(x)} \right| \leq \text{osc}(H) \cdot |s - t| \quad \text{for all } s, t \geq 0 \text{ and } x \in S \quad (3)$$

where $\text{osc}(H) := \max_S H - \min_S H$. On the other hand, other interpolations, e.g. by a spatial coarse graining, may be preferable in concrete applications.

One way to obtain sequential methods for Monte Carlo estimation of expectation values with respect to the measures μ_t is to proceed as follows:

- a) Construct a semigroup $(\Phi_{s,t})_{0 \leq s \leq t < \infty}$ of nonlinear transformations on the space of probability measures on S , such that

$$\Phi_{s,t} \mu_s = \mu_t \quad \text{for all } 0 \leq s \leq t. \quad (4)$$

- b) *Spatial discretization by interacting particle system*: Construct an appropriate Markov process (X_t^1, \dots, X_t^N) on S^N ($N \in \mathbb{N}$) related to the nonlinear semigroup $\Phi_{s,t}$, and estimate $\mu_t = \Phi_{0,t} \mu$ by the empirical distributions

$$\hat{\mu}_t^{(N)} := \frac{1}{N} \sum_{i=1}^N \delta_{X_t^i}, \quad t \geq 0,$$

of the process with initial distribution μ^N .

- c) *Time-discretization*: Approximate the continuous time Markov process (X_t^1, \dots, X_t^N) by a time-discrete Markov chain on S^N (which can then be simulated).

1.2. The non-linear semigroup. To define the nonlinear semigroup $\Phi_{s,t}$ and the particle system we have in mind, we consider the generators (Q -matrices) \mathcal{L}_t at time $t \geq 0$ of a time-inhomogeneous Markov chain on S satisfying the detailed balance condition

$$\mu_t(x) \mathcal{L}_t(x, y) = \mu_t(y) \mathcal{L}_t(y, x) \quad \forall t \geq 0, x, y \in S. \quad (5)$$

The generators \mathcal{L}_t determine the MCMC steps in a corresponding sequential MCMC method. We assume that $\mathcal{L}_t(x, y)$ depends continuously on t . To compare algorithmic performance in a reasonable way, one might also assume

$$\sum_{y \neq x} \mathcal{L}_t(x, y) \leq 1 \quad \forall x \in S, \quad (6)$$

although this is not necessary for the results below. For example, \mathcal{L}_t could be the generator of a Metropolis dynamics w.r.t. μ_t , i.e.,

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

$\mathcal{L}_t(x, x) = -\sum_{y \neq x} \mathcal{L}_t(x, y)$, where the proposal matrix K_t is a given symmetric transition matrix on S . By (5), \mathcal{L}_t defines a symmetric linear operator on $L^2(S, \mu_t)$. The associated Dirichlet form on functions $f, g : S \rightarrow \mathbb{R}$ 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 $\mathcal{E}_t(f) := \mathcal{E}_t(f, f)$.

We fix non-negative constants M_t ($t \geq 0$) that determine the average relative frequency of MCMC moves compared to importance sampling/resampling steps in a corresponding SMCMC method. Again, we assume that $t \mapsto M_t$ is continuous.

Let $p_{s,t}(x, y)$ and $q_{s,t}(x, y)$ ($x, y \in S$) be the unique solutions of the forward equations

$$\frac{\partial}{\partial t} p_{s,t} f = p_{s,t} (M_t \mathcal{L}_t f - H f), \quad p_{s,s} f = f, \quad (7)$$

$$\frac{\partial}{\partial t} q_{s,t} f = q_{s,t} (M_t \mathcal{L}_t f - H_t f), \quad q_{s,s} f = f, \quad (8)$$

where

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

The linear semigroups $p_{s,t}$ and $q_{s,t}$ have the Feynman-Kac representations

$$\begin{aligned} p_{s,t} f(x) &= \int e^{-\int_s^t H(X_r(\omega)) dr} f(X_t(\omega)) \mathbb{P}_{t,x}(d\omega), \\ q_{s,t} f(x) &= \int e^{-\int_s^t H_r(X_r(\omega)) dr} f(X_t(\omega)) \mathbb{P}_{t,x}(d\omega), \end{aligned}$$

where $(X_t, \mathbb{P}_{t,x})$ is a time-inhomogeneous Markov process with generator $M_t \cdot \mathcal{L}_t$. In particular one has

$$q_{s,t} f = \exp\left(\int_s^t \mathbb{E}_r[H] dr\right) p_{s,t} f.$$

We consider the nonlinear semigroup

$$\Phi_{s,t} \nu := \nu \frac{\nu p_{s,t}}{(\nu p_{s,t})(S)} = \frac{\nu q_{s,t}}{(\nu q_{s,t})(S)}, \quad 0 \leq s \leq t,$$

on the space $\mathcal{M}_1(S)$ of probability measures on S . Here

$$\nu p(y) = \sum_{x \in S} \nu(x) p(x, y).$$

The semigroup $\Phi_{s,t}$ describes the time evolution of the law of an inhomogeneous Markov chain with generator $M_t \cdot \mathcal{L}_t$ and absorption rate H , conditioned to be alive at time t (see e.g. [2]). It is not difficult to verify that (4) holds, cf. Theorem 1 below.

1.3. Particle system approximations. To approximate $\Phi_{s,t}$ one could use a particle system consisting of independent Markov chains with absorption, and base the Monte Carlo estimation on the particles that are still alive at time t . However, such a procedure would be usually very inefficient, since in most interesting cases the overwhelming majority of particles would have become extinct already at the final time. Instead, sequential Monte Carlo samplers are based on a time-inhomogeneous Markov chain on S^N , $N \in \mathbb{N}$, with a generator that is for example of type

$$\begin{aligned} \bar{\mathcal{L}}_t^N f(x^1, \dots, x^N) = & M_t \cdot \sum_{i=1}^N (\mathcal{L}_t^{(i)} f)(x^1, \dots, x^N) \\ & + \frac{1}{N} \sum_{i,j=1}^N H(x^i) \cdot (f(r^{i,j}(x)) - f(x)). \end{aligned}$$

Here $\mathcal{L}_t^{(i)}$ denotes the application of \mathcal{L}_t to the i -th component, and $r^{i,j}(x) := y$ where $y^i := x^j$ and $y^k := x^k$ for all $k \neq i$. Hence, between the interactions the particles move according to time-inhomogeneous Markov chains with generator $M_t \cdot \mathcal{L}_t$ and absorption rate H , and in case of absorption, the position is replaced by the position of a randomly chosen particle. Other interaction terms that correspond to different resampling schemes are possible as well. The asymptotics as $N \rightarrow \infty$ of the approximating particle systems with mean field interaction has been studied intensively, cf. e.g. the monograph [4].

1.4. Convergence and stability properties. The quality of Monte Carlo estimates of $\mu_t(f) = \int f d\mu_t$ for some function $f : S \rightarrow \mathbb{R}$ can be measured by the bias and the (asymptotic) variance of the corresponding estimators. The theoretical analysis of the sequential MCMC methods considered here can be subdivided into several steps as above:

- a) Stability properties of the semigroup $\Phi_{s,t}$.
- b) Bias and asymptotic variance of the estimators $\hat{\mu}_t^N(f) = \frac{1}{N} \sum_{i=1}^N f(X_t^i)$.
- c) Effect of the discretization in time.

In this paper, we will focus exclusively on the first step, that is we develop a stability analysis for $\Phi_{s,t}$ based on functional inequalities. A follow-up paper [13] will be devoted to the time dependence of the asymptotic (as $N \rightarrow \infty$) mean square error of the particle

system based estimators $\hat{\mu}_t^N(f)$. Let us remark for the moment, that significant work in this direction has already been done, e.g., by Del Moral and Miclo in [8]. The results clearly indicate that techniques very close to those developed here can also be applied to control the asymptotic variances of the approximating particle systems. This will be made precise in [13].

We also point out that usually the time discretization is carried out before the spatial discretization, i.e. one usually directly considers semigroups and particle systems in discrete time. Even though this is closer to the algorithmic realization, the convergence analysis becomes more transparent in continuous time due to the infinitesimal description (at least from an analytic perspective). Moreover, the continuous time setup allows us to see more clearly how frequently different types of moves of the particle systems should be carried out.

Before stating our results, we comment on relations of sequential MCMC methods to several standard methods for Monte Carlo integration:

- *Parallel MCMC* is a special case of the algorithm above when $H \equiv 0$, i.e. $\mu_t = \mu$ for all t . In this case the associated particle system consists of independent *time-homogeneous* Markov chains with invariant measure μ . Common problems are slow mixing due to multimodality and the burn-in time (i.e. the time needed to reach equilibrium from an initial distribution that is far from μ can be much larger than the inverse spectral gap). Both problems are particularly significant in high dimensional setups (“curse of dimension”).

- In *parallel simulated annealing*, the approximating particle system is given by independent time-inhomogeneous Markov chains with generator \mathcal{L}_t . There are no interactions. In this case, the corresponding (linear) semigroup on probability measures does not satisfy (4). As a consequence, there is an asymptotic bias of the corresponding Monte Carlo estimator, which can only be reduced by the mixing properties of the underlying Markov chains. Therefore, in multimodal setups good convergence properties can only be guaranteed if the measures μ_t change very slowly (logarithmic cooling schedule).

- Pure *importance sampling/resampling* is the special case of our method when $\mathcal{L}_t \equiv 0$ for all t . Since the particles cannot explore the state space, it is only applicable for small state spaces, or in very special situations. In fact, the results below indicate that a certain amount of particle motion is needed to ensure good stability properties. Our results below can be used to quantify, at least in principle, how many MCMC moves are needed to balance the error growth due to importance sampling/resampling.

– A combination of importance sampling and MCMC (without resampling) is similar to considering Markov chains with absorption, conditioned to stay alive. This is often inefficient, cf. the remark above.

– Finally, we would like to point out that the analysis of several multilevel sampling methods (see e.g. [15]) such as *umbrella sampling* (cf. [16]), *simulated* and *parallel tempering* (cf. [18], [1], [21]) has been an inspiration for this work. These MCMC methods provide samples from mixtures, direct sums, or products of distributions μ_{t_i} ($0 \leq i \leq m$), $0 = t_0 < t_1 < \dots < t_m$, $m \in \mathbb{N}$. A disadvantage of umbrella sampling and simulated tempering is that the normalization constants have to be estimated in parallel. Parallel tempering avoids this disadvantage by simulating the product distribution $\bigotimes_{i=0}^m \mu_{t_i}$ with the help of a Metropolis chain on S^{m+1} where neighboring coordinates can be swapped. However, the swapping procedure seems to slow down the convergence in some cases, and it makes the convergence analysis rather intricate, cf. [18]. The sequential MCMC methods presented here can be seen as an attempt to overcome these difficulties. The estimation of the normalization constant is built into the algorithm, and the evolution in t is linear – and thus faster than a diffusive motion in t as in simulated and parallel tempering. Once the basic techniques are developed, the convergence analysis seems also to be at least partially more transparent for sequential MCMC than for simulated and parallel tempering.

2. MAIN RESULTS

2.1. Time evolution of the mean square error. Let $\nu_t := \Phi_{0,t}\nu$ for some given initial distribution $\nu \in \mathcal{M}_1(S)$, and let

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

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

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

denote the mean square error (χ^2 -contrast) of ν_t w.r.t. μ_t . Our first result shows that $\mu_t = \Phi_{0,t}\mu_0$, and it gives a general method to analyze the stability of this evolution in an L^2 sense:

Theorem 1. (i) $\nu_t = \Phi_{0,t}\nu$ is the unique solution of the nonlinear evolution equation

$$\frac{\partial}{\partial t} \nu_t = M_t \nu_t \mathcal{L}_t - H \nu_t + \nu_t(H) \nu_t, \quad t \geq 0. \quad (9)$$

with initial condition $\nu_0 = \nu$.

(ii) The densities g_t solve

$$\frac{\partial}{\partial t} g_t = M_t \mathcal{L}_t g_t + \mathbb{E}_t[H(g_t - 1)] g_t. \quad (10)$$

(iii) The time evolution of the mean square error is given by

$$\frac{1}{2} \frac{d}{dt} \varepsilon_t = -M_t \mathcal{E}_t(g_t - 1) - \frac{1}{2} \mathbb{E}_t[H_t(g_t - 1)^2] + \mathbb{E}_t[H_t(g_t - 1)] \varepsilon_t. \quad (11)$$

Remark 2. (9) is the forward equation for the nonlinear semigroup $\Phi_{s,t}$ (for $s = 0$). The corresponding assertion holds for $\nu_t := \Phi_{s,t}\nu$ for $t \geq s > 0$. Since μ_t solves (9), we obtain in particular

$$\mu_t = \Phi_{s,t}\mu_s \quad \text{for all } t \geq s \geq 0.$$

The proof of the theorem is given in Section 3 below. Similar equations have been derived in a more general setup by Stannat [22].

The main objective of this article is to develop efficient tools to bound the growth of ε_t based on Theorem 1. To estimate the right-hand side of (11) 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.

2.2. Stability based on global estimates. For $t \geq 0$ let

$$C_t := \sup \{ \mathbb{E}_t[f^2] / \mathcal{E}_t(f, f) \mid f : S \rightarrow \mathbb{R} \text{ s.t. } \mathbb{E}_t[f] = 0, f \neq 0 \}$$

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

$$A_t := \sup \{ \mathbb{E}_t[H_t^- f^2] / \mathcal{E}_t(f, f) \mid f : S \rightarrow \mathbb{R} \text{ s.t. } \mathbb{E}_t[f] = 0, f \neq 0 \}.$$

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

$$\text{Var}_t(f) \leq C_t \cdot \mathcal{E}_t(f, f) \quad \forall f : S \rightarrow \mathbb{R}, \quad \text{and} \quad (12)$$

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

Here Var_t denotes the variance w.r.t. μ_t .

Remark 3. (i) There exist efficient techniques to obtain upper bounds for C_t , for example the method of canonical paths, comparison methods (see e.g. [20]), as well as decomposition methods (see e.g. [14]). Variants of these techniques can be applied to estimate A_t as well.

(ii) Clearly, one has

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

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

Let

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

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 4. *If $M_t \geq A_t/2$ for all $t \geq 0$, then*

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

and

$$\frac{d}{dt} \log \varepsilon_t \leq -\frac{2M_t - A_t}{C_t} + 2\left(\frac{A_t}{C_t} \mathbb{E}_t[H_t^-]\right)^{1/2} \varepsilon_t^{1/2} + \mathbb{E}_t[H_t^-] \varepsilon_t. \quad (16)$$

The proof will be given in Section 3 below. Inequality (15) is straightforward to prove, but sometimes (16) is stronger, since the constants only depend on the negative part of H_t . As an immediate consequence of the theorem we obtain estimates on the average relative frequency M_t of MCMC moves that is sufficient to guarantee stability of the corresponding nonlinear flow of probability measures:

Corollary 5. *Let $0 \leq \beta_0 < \beta_1$, and assume that for all $t \in (\beta_0, \beta_1)$,*

$$M_t > \frac{A_t}{2} + C_t \sigma_t(H) \varepsilon_{\beta_0}^{1/2} \quad (17)$$

or

$$M_t > \frac{A_t}{2} + (A_t C_t \mathbb{E}_t[H_t^-])^{1/2} \varepsilon_{\beta_0}^{1/2} + \frac{1}{2} C_t \mathbb{E}_t[H_t^-] \varepsilon_{\beta_0}. \quad (18)$$

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

Remark 6. (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 given below.

(ii) Similarly to the 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

$$M_t > \frac{A_t + \gamma C_t}{2} + C_t \sigma_t(H) e^{-\gamma(t-\beta_0)/2} \varepsilon_{\beta_0}^{1/2} \quad \forall t \in (\beta_0, \beta_1), \quad (19)$$

or a similar condition replacing (18) holds.

(iii) One can often assume that the initial error ε_{β_0} is very small. In this case, M_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 $\partial \varepsilon_t / \partial t \leq -2 \frac{M_t}{C_t} \varepsilon_t$. This yields the classical exponential decay with rate 2γ of the mean square error in the presence of the global spectral gap $M_t/C_t \geq \gamma$ of the generator $M_t \cdot \mathcal{L}_t$. For $H \neq 0$, additional MCMC moves are required to make up for the error growth due to importance sampling/resampling.

Roughly, the corollary says that if 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 $M_t \geq A_t/2$.

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}_β . 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

$$\varepsilon_t \leq e^{-2t/C_\beta} \varepsilon_0 \leq e^{-2t/C_\beta} \cdot (e^{\beta \operatorname{osc}(H)} - 1)$$

where we have used that

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

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

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

is sufficient. If (6) 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 (18), we see that ε_t decreases in time if

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

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.

2.3. Error control based on local estimates. Madras and Randall [17] and Jerrum, Son, Tetali and Vigoda [14] 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 [18], [1] and [21]. 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 $0 \leq \beta_0 < \beta_1 \leq \infty$. 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

$$\sum_i \mu_t(S_t^i) \mathcal{E}_t^i(f, f) \leq K \cdot \mathcal{E}_t(f, f) \quad \forall t \in (\beta_0, \beta_1), f : S \rightarrow \mathbb{R} \quad (20)$$

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.,

$$\mathcal{E}_t^i(f, f) = \frac{1}{2} \sum_{x, y \in S_t^i} (f(y) - f(x))^2 \mathcal{L}_t(x, y) \mu_t(x | S_t^i). \quad (21)$$

In this case, (20) 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 $\pi : S \rightarrow I$ 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 $f : S \rightarrow \mathbb{R}$. We set

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

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

$$\mathbb{E}_t^i[-\tilde{H}_t f^2] \leq A_t^i \cdot \mathcal{E}_t^i(f, f) \quad \forall f : S \rightarrow \mathbb{R} : \mathbb{E}_t[f | \pi] = 0, \quad (22)$$

$$|\mathbb{E}_t^i[\tilde{H}_t f]|^2 \leq B_t^i \cdot \mathcal{E}_t^i(f, f) \quad \forall f : S \rightarrow \mathbb{R} : \mathbb{E}_t[f | \pi] = 0. \quad (23)$$

Remark 7. (i) Note that to verify (22) it is enough to estimate $\mathbb{E}_t^i[\tilde{H}_t^- f^2]$, while for (23) one has to take into account the positive part of \tilde{H}_t as well. In particular, (22) can not be used to derive an estimate of type (23). However, if (22) holds with $-\tilde{H}_t$ replaced by $|\tilde{H}_t|$, then (23) holds with $B_t^i = \mathbb{E}_t^i[|\tilde{H}_t|] \cdot A_t^i$.

(ii) If local Poincaré inequalities of the type

$$\text{Var}_t^i(f) \leq C_t^i \cdot \mathcal{E}_t^i(f, f) \quad \forall f : S \rightarrow \mathbb{R}, i \in I, \quad (24)$$

hold, then (22) and (23) hold with $A_t^i = C_t^i \cdot \max_{S_i} \tilde{H}_t^-$ and $B_t^i = C_t^i \cdot \text{Var}_t^i(H)$.

Combining (20) and (22), (23) respectively yields

$$\mathbb{E}_t[-\tilde{H}_t \tilde{f}_t^2] = \sum_{i \in I} \mu_t(S_t^i) \mathbb{E}_t^i[-\tilde{H}_t \tilde{f}_t^2] \leq \hat{A}_t \cdot \mathcal{E}_t(f, f) \quad \forall f : S \rightarrow \mathbb{R}, \quad (25)$$

and

$$\sum_{i \in I} \mu_t(S_t^i) \left| \mathbb{E}_t^i[\tilde{H}_t \tilde{f}_t] \right|^2 \leq \hat{B}_t \cdot \mathcal{E}_t(f, f) \quad \forall f : S \rightarrow \mathbb{R}, \quad (26)$$

where

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

The following error estimate is our key result :

Theorem 8. *If $M_t > \hat{A}_t/2$ for all $t \in (\beta_0, \beta_1)$ then*

$$\frac{d}{dt} \log \varepsilon_t \leq \frac{\hat{B}_t}{M_t - \hat{A}_t/2} \cdot (1 + \varepsilon_t) + (1 + \sqrt{\varepsilon_t})^2 \cdot \max_{i \in I} h_t^-(i) \quad (27)$$

where

$$h_t(i) := \mathbb{E}_t^i[H] - \mathbb{E}_t[H] = - \left. \frac{\partial}{\partial s} \log \mu_s(S_t^i) \right|_{s=t} \quad (i \in I). \quad (28)$$

The proof is given in Section 3 below. To understand the consequences of (27), let us first consider the asymptotics as M_t tends to infinity. In this case, (27) 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

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

Thus

$$\min(\varepsilon_t, \delta) \leq \varepsilon_{\beta_0} \cdot G_t^{(1+\sqrt{\delta})^2} \quad \forall t \in [\beta_0, \beta_1], \quad (29)$$

where

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

Remark 9. The term $G_t^{(1+\sqrt{\delta})^2}$ in (29) 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 $\frac{\partial}{\partial s} \log \mu_s(S^i)$ is maximized by S^{i_0} for all $s \in (\beta_0, \beta_1)$, then

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

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 M_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 (29).

Corollary 10. *Let $\beta \in (\beta_0, \beta_1]$ and $\delta > 0$, and assume that*

$$M_t \geq \frac{\hat{A}_t}{2} + \alpha_t \cdot \hat{B}_t \quad \forall t \in (\beta_0, \beta) \quad (30)$$

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

$$\min(\varepsilon_\beta, \delta) \leq \varepsilon_{\beta_0} \cdot G_\beta^{(1+\sqrt{\delta})^2} \cdot \exp \int_{\beta_0}^\beta \frac{1+\delta}{\alpha_s} ds. \quad \forall t \in [\beta_0, \beta]. \quad (31)$$

In particular, if

$$M_t \geq \frac{\hat{A}_t}{2} + (\beta - \beta_0) \cdot \hat{B}_t \quad \forall t \in (\beta_0, \beta) \quad (32)$$

then

$$\min(\varepsilon_\beta, \delta) \leq \varepsilon_{\beta_0} \cdot G_\beta^{(1+\sqrt{\delta})^2} \cdot e^{1+\delta}. \quad (33)$$

Remark 11. The main difference to Corollary 5 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.

2.4. Example 1: Exponential model with k valleys in the energy landscape.

This is an extended version of a model considered in [16], [18] as a test case for some multi-level MCMC methods. We fix $k \in \mathbb{N}$, and $r_1, r_2, \dots, r_k \in \mathbb{N}$. Let $S^0 := \{0\}$ and

$$S^i := \{(i, j) : j = 1, 2, \dots, r_i\}, \quad 1 \leq i \leq k.$$

We consider the graph with vertex set

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

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

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

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 (1), 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.,

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

for all $1 \leq i \leq k$ and $1 \leq j \leq r_i$. 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 $R = \max_{1 \leq i \leq k} r_i$. An application of Corollary 10 with $\beta_0 = 0$ and $S_t^i = S^i$ for all $t \geq 0$ yields the following result :

Theorem 12. *If*

$$M_t \geq R^3 + \frac{\beta}{2} R^4 \quad \forall t \in (0, \beta),$$

then

$$\min(\varepsilon_\beta, \delta) \leq e^{1+\delta} \cdot \varepsilon_0 G_\beta^{(1+\sqrt{\delta})^2} \cdot \varepsilon_0 \quad \forall \delta \in (0, 1). \quad (34)$$

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

$$\mu(\{(i, j) : j \geq h\} | S^{i_0}) \geq \mu(\{(i, j) : j \geq h\} | S^i), \quad (35)$$

then

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

and thus

$$\min(\varepsilon_\beta, \delta) \leq e^{1+\delta} \cdot \frac{\varepsilon_0}{\mu(S^{i_0})^{(1+\sqrt{\delta})^2}} \quad \forall 0 < \delta < 1. \quad (36)$$

Remark 13. (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 $K_\beta = \int_0^\beta M_t dt$. 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 M_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 .

2.5. Example 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 (1) where $\mu_0 = \mu$ is the uniform distribution on the hypercube

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

and

$$H(\sigma) = -\frac{1}{2N} \sum_{i,j=1}^N \sigma_i \sigma_j \quad (37)$$

for some $N \in \mathbb{N}$. Let \mathcal{L}_β 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

$$\begin{aligned} S^+ &:= \left\{ \sigma \in S \mid \sum_{i=1}^N \sigma_i > 0 \right\} \quad \text{and} \\ S^- &:= \left\{ \sigma \in S \mid \sum_{i=1}^N \sigma_i < 0 \right\}. \end{aligned}$$

Improving on previous results (e.g. of Madras and Zheng [18]), Schweizer [21] 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 14. *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

$$M_t \geq \frac{9}{4} N^3 + \frac{9}{8} \beta N^4 \quad \forall t \in (0, \beta). \quad (38)$$

Remark 15. (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 (38). 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 [16], and for simulated and parallel tempering in [18], [1], [21]. Schweizer [21] 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 (38). 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.

2.6. Extensions. As remarked above, our results immediately extend to the case where

$$\mu_t(x) = \frac{1}{Z_t} e^{-\int_0^t U_s(x) ds} \mu(x) \quad (0 \leq t \leq \beta)$$

for a continuous function $(s, x) \mapsto U_s(x)$ on $[0, \beta]$ with $U_s(x) \geq 0$ for all $s \in [0, \beta]$ and $x \in S$. In this case, the evolution equations (9) and (10) in Theorem 1 take the form

$$\begin{aligned} \frac{\partial}{\partial t} \nu_t &= M_t \nu_t \mathcal{L}_t - U_t \nu_t + \nu_t(U_t) \nu_t, & \text{and} \\ \frac{\partial}{\partial t} g_t &= M_t \mathcal{L}_t g_t + \mathbb{E}_t[U_t(g_t - 1)] g_t. \end{aligned}$$

The evolution equation (11) for the mean square error and all the stability estimates in Sections 2.2 and 2.3 still hold if H is replaced by U_t , and, correspondingly,

$$H_t = U_t - \mathbb{E}_t[U_t].$$

All proofs are completely analogous.

The extension of the results to more general state spaces requires some (standard) technical assumptions which make the proofs slightly less transparent. We postpone this extension to a future publication where we will also consider corresponding applications.

3. PROOFS

3.1. Proof of Theorem 1. To simplify the notation, we assume $M_t = 1$ for all $t \geq 0$. The general case is similar with \mathcal{L}_t replaced by $M_t \cdot \mathcal{L}_t$. Let us also set $p_t := p_{0,t}$ and $\Phi_t := \Phi_{0,t}$. Then one has

$$\nu_t = \Phi_t \nu = \frac{\nu p_t}{(\nu p_t)(1)}.$$

The forward equation (7) yields

$$\frac{\partial}{\partial t} \nu p_t = \nu p_t \mathcal{L}_t - H \nu p_t.$$

Since $(\nu p_t)(1) > 0$ and $\mathcal{L}_t 1 = 0$ for all t , we obtain

$$\begin{aligned} \frac{\partial}{\partial t} \nu_t &= \frac{\partial}{\partial t} \frac{\nu p_t}{(\nu p_t)(1)} \\ &= \frac{\nu p_t \mathcal{L}_t - H \nu p_t}{(\nu p_t)(1)} - \frac{(\nu p_t \mathcal{L}_t - H \nu p_t)(1) \nu p_t}{(\nu p_t)(1)^2} \\ &= \nu_t \mathcal{L}_t - H \nu_t + \nu_t(H) \nu_t. \end{aligned} \tag{39}$$

$$= \nu_t \mathcal{L}_t - H \nu_t + \nu_t(H) \nu_t. \tag{40}$$

Next, we derive a corresponding evolution equation for the densities

$$g_t(y) := \frac{\nu_t(y)}{\mu_t(y)} \quad (y \in S).$$

Since μ_t has full support and is differentiable in t , we obtain

$$\frac{\partial}{\partial t} g_t = \frac{1}{\mu_t} \frac{\partial}{\partial t} \nu_t - \frac{\nu_t}{\mu_t} \frac{\partial}{\partial t} \log \mu_t. \tag{41}$$

Note that by the detailed balance condition (5), the relative density of $\nu_t \mathcal{L}_t$ w.r.t. μ_t is

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

Hence (40) yields

$$\begin{aligned} \frac{1}{\mu_t} \frac{\partial}{\partial t} \nu_t &= \mathcal{L}_t g_t - H g_t + \nu_t(H) g_t \\ &= (\mathcal{L}_t - H) g_t + \mathbb{E}_t[H g_t] g_t. \end{aligned} \tag{42}$$

Recalling that $\mu_t = \frac{1}{Z_t} e^{-tH} \mu$ with $Z_t = \sum_S e^{-tH(y)} \mu(y)$, one has

$$\frac{\partial}{\partial t} \log \mu_t = -\mu_t (H - \mathbb{E}_t[H]) = -\mu_t H_t, \tag{43}$$

hence

$$-\frac{\nu_t}{\mu_t} \frac{\partial}{\partial t} \log \mu_t = (H - \mathbb{E}_t[H]) g_t. \tag{44}$$

Inserting (42) and (44) into (41) we obtain

$$\begin{aligned}\frac{\partial}{\partial t}g_t &= \mathcal{L}_t g_t + (\mathbb{E}_t[Hg_t] - \mathbb{E}_t[H])g_t \\ &= \mathcal{L}_t g_t + \mathbb{E}_t[H(g_t - 1)]g_t.\end{aligned}\tag{45}$$

We are now ready to derive the equation for the quadratic error

$$\varepsilon_t = \mathbb{E}_t[(g_t - 1)^2].$$

Differentiating this expression with respect to t one gets by (45) and (43),

$$\begin{aligned}\frac{d}{dt}\varepsilon_t &= 2\mathbb{E}_t\left[\left(\frac{\partial}{\partial t}g_t\right)(g_t - 1)\right] + \mathbb{E}_t\left[(g_t - 1)^2\frac{\partial}{\partial t}\log\mu_t\right] \\ &= 2\mathbb{E}_t[(\mathcal{L}_t g_t)(g_t - 1)] + 2\mathbb{E}_t[g_t(g_t - 1)]\mathbb{E}_t[H(g_t - 1)] - \mathbb{E}_t[H_t(g_t - 1)^2] \\ &= 2\mathbb{E}_t[\mathcal{L}_t(g_t - 1)(g_t - 1)] + 2\mathbb{E}_t[(g_t - 1)^2]\mathbb{E}_t[H(g_t - 1)] - \mathbb{E}_t[H_t(g_t - 1)^2] \\ &= -2\mathcal{E}_t(g_t - 1) + 2\mathbb{E}_t[H(g_t - 1)]\cdot\varepsilon_t - \mathbb{E}_t[H_t(g_t - 1)^2].\end{aligned}$$

In the derivation we have used that

$$\mathbb{E}_t[g_t - 1] = \nu_t(1) - \mu_t(1) = 0,$$

and $\mathcal{L}_t 1 \equiv 0$. The equation implies (11) in the case $M_t \equiv 1$. The general case follows similarly. \square

3.2. Proof of Theorem 4. We have to estimate the terms on the right hand side of (11). By the assumed H -Poincaré inequality (13), we obtain

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

Moreover,

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

Substituting into (11) yields

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

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

On the other hand,

$$\begin{aligned}
& \mathbb{E}_t \left[H_t \left(- (g_t - 1)^2 / 2 + (g_t - 1) \varepsilon_t \right) \right] \\
&= \frac{1}{2} \mathbb{E}_t \left[H_t^- (g_t - 1)^2 \right] + \mathbb{E}_t \left[H_t^- (1 - g_t) \right] \varepsilon_t \\
&\quad + \mathbb{E}_t \left[H_t^+ \left(- (g_t - 1)^2 / 2 + (g_t - 1) \varepsilon_t \right) \right].
\end{aligned} \tag{46}$$

Estimating the three summands on the right hand side separately yields

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

by the H -Poincaré inequality (13),

$$\begin{aligned}
\mathbb{E}_t \left[H_t^- (1 - g_t) \right] &\leq \mathbb{E}_t \left[H_t^- \right]^{1/2} \mathbb{E}_t \left[H_t^- (g_t - 1)^2 \right]^{1/2} \\
&\leq \mathbb{E}_t \left[H_t^- \right]^{1/2} A_t^{1/2} \mathcal{E}_t(g_t - 1)^{1/2}
\end{aligned}$$

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

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

The last estimate follows since

$$\frac{1}{2} \varepsilon_t^2 \geq (g_t - 1) \varepsilon_t - \frac{1}{2} (g_t - 1)^2$$

and

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

By combining the estimates, (46) and (11), we obtain

$$\frac{d}{dt} \varepsilon_t \leq -(2M_t - A_t) \mathcal{E}_t(g_t - 1) + 2A_t^{1/2} \mathbb{E}_t \left[H_t^- \right]^{1/2} \mathcal{E}_t(g_t - 1)^{1/2} \varepsilon_t + \mathbb{E}_t \left[H_t^- \right] \varepsilon_t^2.$$

This combined with the global Poincaré inequality (12) yields

$$\frac{d}{dt} \varepsilon_t \leq -\frac{2M_t - A_t}{C_t} \varepsilon_t + 2 \frac{A_t^{1/2}}{C_t^{1/2}} \mathbb{E}_t \left[H_t^- \right]^{1/2} \varepsilon_t^{3/2} + \mathbb{E}_t \left[H_t^- \right] \varepsilon_t^2,$$

and hence (16). □

3.3. Proof of Corollary 5. If (17) or (18) holds for $t \in (\beta_0, \beta_1)$, then by Theorem 4 and continuity, $t \mapsto \varepsilon_t$ is strictly decreasing near β_0 and near any $s \in (\beta_0, \beta_1)$ such that $\varepsilon_s \leq \varepsilon_{\beta_0}$. Hence it is strictly decreasing on the whole interval $[\beta_0, \beta_1]$. □

3.4. Proof of Theorem 8. Similarly to Theorem 4, we have to control the right hand side of (11), but now by using only local Poincaré type inequalities. Let

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

Then

$$\begin{aligned} & \mathbb{E}_t [H_t \cdot (-(g_t - 1)^2/2 + (g_t - 1)\varepsilon_t)] \\ &= \mathbb{E}_t [\tilde{H}_t (-(g_t - 1)^2/2 + (g_t - 1)\varepsilon_t)] \\ & \quad + \sum_{i \in I} \mu_t(S_t^i) (\mathbb{E}_t^i[H] - \mathbb{E}_t[H]) \cdot \mathbb{E}_t^i [-(g_t - 1)^2/2 + (g_t - 1)\varepsilon_t] \\ &= -\frac{1}{2} \mathbb{E}_t[\tilde{H}_t \tilde{f}_t^2] - \mathbb{E}_t[\tilde{H}_t \tilde{f}_t \mathbb{E}_t[f_t|\pi]] + \mathbb{E}_t[\tilde{H}_t \tilde{f}_t \varepsilon_t] \\ & \quad + \sum_{i \in I} \mu_t(S_t^i) \cdot (\mathbb{E}_t^i[H] - \mathbb{E}_t[H]) \cdot \mathbb{E}_t^i [-f_t^2/2 + f_t \varepsilon_t] \tag{47} \\ &= -\frac{1}{2} \mathbb{E}_t[\tilde{H}_t \tilde{f}_t^2] + \sum_{i \in I} \mu_t(S_t^i) \cdot \mathbb{E}_t^i[\tilde{H}_t \tilde{f}_t] \cdot (\varepsilon_t - \mathbb{E}_t^i[f_t]) \\ & \quad + \sum_{i \in I} \mu_t(S_t^i) h_t(i) \cdot \mathbb{E}_t^i[-f_t^2/2 + f_t \varepsilon_t]. \end{aligned}$$

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 (25),

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

By (26), and since

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

we have

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

Moreover, since

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

we obtain

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

Here we have used that

$$\begin{aligned}
\sum \mu_t(S_t^i) h_t^+(i) &= \sum \mu_t(S_t^i) h_t^-(i) \leq \max h_t^-, \quad \text{and} \\
\sum \mu_t(S_t^i) \mathbb{E}_t^i[-f_t] &\leq \left(\sum \mu_t(S_t^i) \mathbb{E}_t^i[f_t^2] \right)^{1/2} = \varepsilon_t^{1/2}.
\end{aligned}$$

Combining the estimates yields by (11) and (47) :

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

provided $M_t > \hat{A}_t/2$. This proves (27).

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

$$\begin{aligned}
\frac{d}{dt} \log \mu_t(A) &= \frac{d}{dt} \log \sum_{x \in A} e^{-tH(x)} \mu(x) - \frac{d}{dt} \log Z_t \\
&= -\mathbb{E}_t[H|A] + \mathbb{E}_t[H],
\end{aligned}$$

which proves (28). □

3.5. Proof of Corollary 10. Assume that (30) holds, and let

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

Then by the definition of G_t , Theorem 8, and (30),

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

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

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

holds for $t \in [\beta_0, \beta]$ provided the right hand side is smaller than δ . This proves (31).

The second assertion is a straightforward consequence. \square

3.6. Proof of Theorem 12. 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 [12]. Now we are in the setting of Remark 7 (ii), according to which (22) and (23) hold with \mathcal{E}_t^i as in (21),

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

Estimate (34) now follows by a straightforward application of Corollary 10.

To prove the second part of the assertion, we show that (35) places us in the setting of Remark 9. In fact, for $t > 0$,

$$\begin{aligned} \frac{d}{dt} \log \mu_t(S^i) &= \mathbb{E}_t[H] - \mathbb{E}_t^i[H] \quad \text{for all } i, \text{ and} \\ -\mathbb{E}_t^i[H] &= -\frac{\mu(H e^{-tH} | S^i)}{\mu(e^{-tH} | S^i)} = \frac{\sum_j j e^{tj} \mu((i, j))}{\sum_j e^{tj} \mu((i, j))}. \end{aligned}$$

If (35) holds, then for any $t > 0$, the right hand side is maximized when $i = i_0$. Hence by Remark 9,

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

\square

3.7. Proof of Theorem 14. Since $-N/2 \leq H(\sigma) \leq 0$ for all σ , we have $\text{osc}(H) \leq N/2$ and

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

for every $t \geq 0$. By Schweizer's result [21], a local Poincaré inequality of type (24) holds on S^+ and S^- with $C_t^+ = C_t^- = 9N^2$. Hence by Remark 7 (ii), (22) and (23) hold with

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

The assertion now follows from Corollary 10, since

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



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. R. M. Blumenthal and R. K. Gettoor, *Markov processes and potential theory*, Pure and Applied Mathematics, Vol. 29, Academic Press, New York, 1968. MR MR0264757 (41 #9348)
3. O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*, Springer Series in Statistics, Springer, New York, 2005. MR MR2159833 (2006e:60002)
4. P. Del Moral, *Feynman-Kac formulae*, Springer-Verlag, New York, 2004. MR MR2044973 (2005f:60003)
5. P. Del Moral and A. Doucet, *On a class of genealogical and interacting Metropolis models*, Séminaire de Probabilités XXXVII, Lecture Notes in Math., vol. 1832, Springer, Berlin, 2003, pp. 415–446. MR MR2053058 (2005g:65013)
6. P. Del Moral, A. Doucet, and A. Jasra, *Sequential Monte Carlo samplers*, J. R. Statist. Soc. B **68** (2006), no. 3, 411–436. MR MR1819122 (2002k:60013)
7. P. Del Moral and A. Guionnet, *On the stability of interacting processes with applications to filtering and genetic algorithms*, Ann. Inst. H. Poincaré Probab. Statist. **37** (2001), no. 2, 155–194. MR MR1819122 (2002k:60013)
8. P. Del Moral and L. Miclo, *Branching and interacting particle systems approximations of Feynman-Kac formulae with applications to non-linear filtering*, Séminaire de Probabilités, XXXIV, Lecture Notes in Math., vol. 1729, Springer, Berlin, 2000, pp. 1–145. MR MR1768060 (2001g:60091)
9. 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)
10. ———, *Logarithmic Sobolev inequalities for finite Markov chains*, Ann. Appl. Probab. **6** (1996), no. 3, 695–750. MR MR1410112 (97k:60176)
11. ———, *Nash inequalities for finite Markov chains*, J. Theoret. Probab. **9** (1996), no. 2, 459–510. MR MR1385408 (97d:60114)
12. ———, *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)
13. A. Eberle and C. Marinelli, *Convergence of sequential Markov chain Monte Carlo methods: II. Asymptotic analysis of interacting particle systems*, In preparation.
14. 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)
15. J. S. Liu, *Monte Carlo strategies in scientific computing*, Springer-Verlag, New York, 2001. MR MR1842342 (2002i:65006)

16. N. Madras and M. Piccioni, *Importance sampling for families of distributions*, Ann. Appl. Probab. **9** (1999), no. 4, 1202–1225. MR MR1728560 (2001e:60139)
17. 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)
18. N. Madras and Z. Zheng, *On the swapping algorithm*, Random Structures Algorithms **22** (2003), no. 1, 66–97. MR MR1943860 (2004c:82117)
19. C. P. Robert and G. Casella, *Monte Carlo statistical methods*, second ed., Springer-Verlag, New York, 2004. MR MR2080278 (2005d:62006)
20. 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)
21. N. Schweizer, *Diploma thesis, Universität Bonn*, 2006.
22. W. Stannat, *On the convergence of genetic algorithms—a variational approach*, Probab. Theory Related Fields **129** (2004), no. 1, 113–132. MR MR2052865 (2005d:35040)

INSTITUT FÜR ANGEWANDTE MATHEMATIK, UNIVERSITÄT BONN, WEGELERSTR. 6, 53115 BONN,
GERMANY

E-mail address: eberle@uni-bonn.de

URL: <http://wiener.iam.uni-bonn.de/~eberle>

INSTITUT FÜR ANGEWANDTE MATHEMATIK, UNIVERSITÄT BONN, WEGELERSTR. 6, 53115 BONN,
GERMANY

URL: <http://wiener.iam.uni-bonn.de/~marinelli>