RELAXATION OF PRODUCT MARKOV CHAINS ON PRODUCT SPACES

PETER MATHÉ

Date: January 28, 1997.

1991 *Mathematics Subject Classification.* 60J15.

Key words and phrases. Product Markov chain, Mixing Time, Metropolis sampler.

ABSTRACT. The purpose of the paper is studying the relaxation time of product–type Markov chains on product spaces which approach a product distribution. We determine bounds to approach stationarity for such Markov chains in terms of the mixing times of the component Markov chains. In cases where the component mixing times vary much we propose an optimized visiting scheme which makes such product–type Markov chains comparative to Gibbs–type samplers.

We conclude the paper by a discussion of the relaxation of Metropolis–type samplers applied to separable energy functions.

1. INTRODUCTION, BACKGROUND

Sampling from given distributions even from a finite population may be laborious. One way to circumvent this is asymptotically sampling using a strategy called Metropolis sampling. We shall study the efficiency of this procedure within the context of distributions given on product structures. Hence we suppose that we are given d finite sets $X_1, \dots X_d$ and corresponding distributions $\pi_1, \dots, \pi_d$. The prototype of this setup is provided by d–dimensional grids on a given domain in $\mathbb{R}^d$ with possibly direction dependent mesh size (suited to a function living on the domain). The purpose of the paper is studying the relaxation time of product–type Markov chains on $X := \prod_{j=1}^d X_j$ which asymptotically approach $\Pi := \prod_{j=1}^d \pi_j$. Of course, this is a serious restriction of the applicability of the results obtained below. Nevertheless we hope pointing at properties required from the given distribution to enable asymptotic sampling without visiting most of the states. Such type of problems will be the subject of Section 5.

A first analysis of this type was carried out within the context of groups in a previous study, [5] by the author. As mentioned there it was not necessary to restrict to the setup of groups and the uniform distribution to be approximated. However the analysis has to be different, since switching from Markov chain to convolution of measures is not possible in the general framework which shall be outlined below.

Suppose we are given Markov chains on the component sets $X_1, \dots X_d$ driven by the respective transition matrices $P_1, \dots, P_d$. A product–type Markov chain is obtained from these components in the following way. We choose a convex combination $\rho := (\rho_1, \dots, \rho_d)$, i.e., $\rho_j \geq 0, \quad \sum_{j=1}^d \rho_j = 1$, and compose

$$P_\rho := \sum_{j=1}^d \rho_j \tilde{P}_j, \tag{1}$$

where ˜ indicates the embedding of the component transition matrices into ones for X. In conjunction with an initial distribution ν on X we obtain a Markov chain on X with respective distribution νP_ρ^n at the n–th step. This corresponds to a mixture of the components and means, that at each step we choose a component of our product space with a certain probability and then we take a transition according to the Markov chain acting on this component. So we may think of ρ as a randomized visiting scheme being the counterpart of the visiting scheme in the context of Gibbs–type samplers, see [7], where this is called a *proposal* or *exploration* distribution.

The mixing behavior of Markov chains shall be quantified in terms of the variation distance of measures. Given a (signed) measure λ on some (finite) set X[1] we denote by

$$\|\lambda\|_X := \max_{A \subset X} |\lambda(A)| = \frac{1}{2} \sum_{x \in X} |\lambda(\{x\})|.$$

Whenever it will be clear from the context, we will suppress the subscript indicating the set the measure is living on. Let us however mention that for a measure λ_j on X_j the corresponding embedded $\tilde{\lambda}_j$ on X obeys $\|\tilde{\lambda}_j\|_X = \|\lambda_j\|_{X_j}$. We also explicitly state an estimate, similar to the one in Lemma (7.9) in [2]:

Let $\bar{P}$ denote any distribution on X.

Lemma 1. *For any distribution P on X which is a mixture*

$$P = \alpha \bar{P} + (1 - \alpha) Q$$

for some choice of $0 < \alpha < 1$ *and distribution Q we have*

$$\bar{P}(\{x, \quad Q(\{x\}) = 0\}) \leq \frac{\|P - \bar{P}\|}{1 - \alpha} \leq 2.$$

Proof. The right–hand side inequality is obvious. To prove the left–hand side estimate let $A := \{x, \quad Q(\{x\}) = 0\}$. On this set A we have $P(\{x\}) = \alpha \bar{P}(\{x\})$ and consequently $P(A^c) = 1 - \alpha \bar{P}(A)$. This implies

$$\begin{aligned} \|P - \bar{P}\| &= \frac{1}{2} \sum_{x \in X} |P(\{x\}) - \bar{P}(\{x\})| \\ &\geq \frac{1}{2} \sum_{x \in A} (1 - \alpha) \bar{P}(\{x\}) + \frac{1}{2} \sum_{x \in A^c} |P(\{x\}) - \bar{P}(\{x\})| \\ &\geq \frac{1}{2}(1 - \alpha) \bar{P}(A) + \frac{1}{2} P(A^c) - \frac{1}{2} \bar{P}(A^c) \\ &\geq (1 - \alpha) \bar{P}(A). \end{aligned}$$

The proof is complete. □

We turn to the study of mixing (relaxation) times. Our approach is close to [1, 2]. Given transition matrices P and Q on X we let

$$d(P, Q) := \max_{\mu \text{ on } X} \|\mu P - \mu Q\| \quad \left(= \max_{x \in X} \|\delta_x P - \delta_x Q\| \right).$$

It is readily seen that this turns to a metric between transition matrices and that with any further transition R we have $d(PR, QR) \leq d(P, Q)$.

Moreover, if μ is a probability on X, then, by letting $P_\mu(x, y) := \mu(\{y\}), \quad x, y \in X$, we agree to write

$$d(P, \mu) := d(P, P_\mu). \tag{2}$$

[1] Throughout the remainder of this section the set X may be arbitrary finite.

In case P is the transition of an ergodic Markov chain with invariant distribution Π we simply abbreviate $d_k(P) := d(P^k, \Pi)$ the (worst) distance of the distribution at the k–th step from the invariant distribution.

As a function of $k \in \mathbb{N}$ it is easily seen to be decreasing. Further, as will be clear below it makes sense to measure the time to reach stationarity in terms of this quantity. So we agree to let

$$K(P) := \min \left\{ k \in \mathbb{N}, \quad d_k(P) \le \frac{1}{2\mathrm{e}} \right\} \tag{3}$$

be the *mixing time* of P. The quantity $d_k(P)$ is close to being submultiplicative. From [5] we recall

Lemma 2. *For any $k \in \mathbb{N}$ the following inequality holds true*

$$d_{l\cdot k}(P) \le (2d_k(P))^l, \quad l \in \mathbb{N}.$$

Especially, with $k := K(P)$ we obtain $d_{l\cdot K(P)}(P) \le \mathrm{e}^{-l}$.

The proof is based on another auxiliary quantity, cf. [1, 2],

$$\rho_k(P) := \max_{x,y\in X} \|\delta_x P^k - \delta_y P^k\|. \tag{4}$$

It is known from Lemma (4.5) in [2] that this is submultiplicative. Moreover we have

$$d_k(P) \le \rho_k(P) \le 2d_k(P). \tag{5}$$

We mention that $\rho_1(P)$ is the contraction coefficient studied in [7, Ch. 4.2], which will be useful in Section 5 below.

In view of Lemma 2 we may think of $K(P)$ as a threshold level starting from which the convergence to stationarity is exponential.

For later use we recall some facts about *multinomial distributions.* Given a d–tuple $\bar{r} = (r_1, \dots, r_d)$ of natural numbers with $r_1 + \dots r_d = k$ we denote by $\binom{k}{\bar{r}} := \frac{k!}{r_1! \cdots r_d!}$ and $r_{min} := \min_{j=1,\dots,d} r_j$. Let $P_{k,\rho}$ denote the multinomial distribution on $\{0, \dots, k\}^d$ with point masses

$$P_{k,\rho}((r_1, \dots, r_d)) = \binom{k}{\bar{r}} \prod_{j=1}^{d} \rho_j^{r_j}, \qquad \text{if } r_1 + \dots r_d = k.$$

A detailed exposition with further references can be found in [4, Ch. 11.2]. We mention that the component distributions of $P_{k,\rho}$ are respective binomial ones B_{k,ρ_j} with respective ρ_j. The following lemma is probably well known and proven in [5].

Lemma 3. *For any d, convex combination ρ and $k \in \mathbb{N}$ we have*

$$1 - \mathrm{e}^{-\sum_{j=1}^{d}(1-\rho_j)^k} \le P_{k,\rho}(\text{"} r_{min} = 0 \text{"}) \le \sum_{j=1}^{d} (1 - \rho_j)^k. \tag{6}$$

2. An auxiliary Markov chain

Below we suppose that we are given d finite state spaces $X_1, \dots, X_d$ with Markov chains driven by respective transition matrices $P_1, \dots, P_d$. Throughout we shall assume that all transition matrices $P_j, \quad j = 1, \dots, d$ are ergodic, hence possess unique invariant distributions denoted by $\pi_1, \dots, \pi_d$, respectively.

A Markov chain on the product $X := \prod_{j=1}^d X_j$ is constructed as follows. We first embed the Markov chains $P_j, \quad j = 1, \dots, d$ into the product by letting for $x = (\xi_1, \dots, \xi_d)$ and $y = (\eta_1, \dots, \eta_d)$ the embedded chain be

$$\tilde{P}_j(x, y) := \begin{cases} P_j(\xi_j, \eta_j) & \text{, if } \xi_l = \eta_l, \ l = 1, \dots, d, \ l \neq j \\ 0 & \text{, otherwise} \end{cases}. \tag{7}$$

Hence, the Markov chains $\tilde{P}_j$ accept transitions in the components X_j only. We mention the following

Lemma 4. *Any 2–step transition $\tilde{P}_i \tilde{P}_j$ is commutative, precisely we have for any $x = (\xi_1, \dots, \xi_d)$ and $y = (\eta_1, \dots, \eta_d)$ the equality*

$$\tilde{P}_i \tilde{P}_j(x, y) = \tilde{P}_j \tilde{P}_i(x, y) = P_i(\xi_i, \eta_i) P_j(\xi_j, \eta_j), \quad i \neq j.$$

For later use we introduce the following

Example. Let $\pi_j, \quad j = 1, \dots, d$, denote the given limit distributions on X_j and consider the Markov chain Q_j, describing an i.i.d. walk on X_j, hence

$$Q_j(\xi_j, \eta_j) := \pi_j(\{\eta_j\}) \quad \xi_j, \eta_j \in X_j, \quad j = 1, \dots, d.$$

Let $\tilde{Q}_j, \quad j = 1, \dots, d$ denote the embeddings of Q_j into X. The following observation is easily checked.

1. The distribution of any 2–step transition $\tilde{Q}_j^2$ equals $\tilde{Q}_j, \quad j = 1, \dots, d$.
2. In view of Lemma 4 we have

$$\tilde{Q}_{i_1} \dots \tilde{Q}_{i_k} = \Pi = \prod_{j=1}^d \pi_j \text{ whenever } \{i_1, \dots, i_k\} = \{1, \dots, d\}.$$

We recall from the introduction that a product–type Markov chain is obtained from these components by choosing a convex combination $\rho := (\rho_1, \dots, \rho_d)$, i.e., $\rho_j \geq 0, \quad \sum_{j=1}^d \rho_j = 1$, and composing

$$P_\rho := \sum_{j=1}^d \rho_j \tilde{P}_j. \tag{8}$$

Especially we shall study the product–type Markov chains Q_ρ obtained from the component transitions $Q_j, \quad j = 1, \dots, d$.

Let us investigate the mixing behavior of the Markov chain Q_ρ introduced before. Recall that the component Markov chains represent i.i.d. samples within the components. For this particular type of walk one can expect that the mixing behavior does not depend on the relaxation times of the involved component Markov chains but rather on the number d of such. This is supported by Lemma 6 below. We need an intermediate fact.

Lemma 5. *Given an initial point $x = (\xi_1, \dots, \xi_d)$ we have for any step number k and $y = (\eta_1, \dots, \eta_d) \in \prod_{j=1}^d (X_j \setminus \{\xi_j\})$ equality*

$$\delta_x Q_\rho^k(\{y\}) = P_{k,\rho}(\text{"} \min\{r_1, \dots, r_d\} > 0\text{"})\Pi(\{y\}), \tag{9}$$

where $(r_1, \dots, r_d)$ counts how many times the respective components have been visited during the k steps.

Proof. Since by Lemma 4 subsequent transitions are commutative we have

$$\begin{aligned}\delta_x Q_\rho^k &= \delta_x \underbrace{\left(\sum_{j=1}^d \rho_j \tilde{Q}_j\right)\left(\sum_{j=1}^d \rho_j \tilde{Q}_j\right) \cdots \left(\sum_{j=1}^d \rho_j \tilde{Q}_j\right)}_{k-\text{fold}} \\ &= \sum_{r_1+\ldots+r_d=k} \binom{k}{\bar{r}} \delta_x \prod_{j=1}^d \left(\rho_j \tilde{Q}_j\right)^{r_j} \quad {}^{2} \\ &= \sum_{\substack{r_1+\ldots+r_d=k \\ r_{min}>0}} \binom{k}{\bar{r}} \prod_{j=1}^d \rho_j^{r_j} \Pi + \sum_{\substack{r_1+\ldots+r_d=k \\ r_{min}=0}} \binom{k}{\bar{r}} \delta_x \prod_{j=1}^d \left(\rho_j \tilde{Q}_j\right)^{r_j}.\end{aligned}$$

For a transition to y which is from $\prod_{j=1}^d (X_j \setminus \{\xi_j\})$ the right–hand side sum above is equal to 0, since for i_0 with $r_{i_0} = 0$ the respective destination η_{i_0} must equal ξ_{i_0} which is impossible by the choice of y. □

This yields

Lemma 6. *For fixed d, convex combination ρ and natural k we have*

$$\prod_{j=1}^d (1 - \frac{1}{|X_j|}) P_{k,\rho}(\text{"} r_{min} = 0\text{"}) \le d_k(Q_\rho) \le 2P_{k,\rho}(\text{"} r_{min} = 0\text{"}). \tag{10}$$

Proof. Choose in each component X_j a point ξ_j^0 with smallest probability $\pi_j(\{\xi_j^0\})$ which is at most $\frac{1}{|X_j|}$. Let this determine the starting point $x_0 := (\xi_1^0, \dots, \xi_d^0)$. In view of equation (9) we apply Lemma 1 to $P := \delta_{x_0} Q_\rho^k$ and $\alpha = P_{k,\rho}(\text{"} r_{min} > 0\text{"})$. Note that by our choice of x_0 we ensure

$$\Pi(\prod_{j=1}^d (X_j \setminus \{\xi_j^0\})) \ge \prod_{j=1}^d (1 - \frac{1}{|X_j|})$$

which completes the proof of the lemma. □

The sharp bounds from Lemma 3 immediately yield

[2] The symbol $\prod$ in conjunction with transition matrices denotes successive transition throughout. Since, in view of Lemma 4, the order does not affect the overall distribution this is justified.

Proposition 1. *If the spaces X_j are rich enough such that $\prod_{j=1}^d (1 - \frac{1}{|X_j|}) \geq \frac{4}{5}$, then we have*

$$K(Q_\rho) \geq d(0.5 + \log(d)).$$

On the other hand, by the choice of $\rho_0 = (\frac{1}{d}, \dots, \frac{1}{d})$ we obtain

$$K(Q_{\rho_0}) \leq d(2.5 + \log(d)).$$

Proof. The proof is an immediate consequence of Lemmas 3 and 6. We only mention that the lower bound in (6) is maximized by letting $\rho = \rho_0 = (\frac{1}{d}, \dots, \frac{1}{d})$. In this case the sum reduces to $d\mathrm{e}^{-k/d}$ and yields with $k = d(0.5 + \log(d))$ the estimate

$$1 - \mathrm{e}^{-\sum_{j=1}^d (1-\rho_j)^k} \geq 1 - \mathrm{e}^{-\mathrm{e}^{-1/2}}.$$

from which the first assertion follows by noting that under our assumptions on X we obtain

$$d_k(Q_\rho) \geq \frac{4}{5}(1 - \mathrm{e}^{-\mathrm{e}^{-1/2}}) \geq \frac{1}{\mathrm{e}}.$$

On the other hand it is easy to see that with $k \geq d(2.5 + \log(d))$ the desired upper bound is obtained, completing the proof of the proposition. □

3. Mixing with fixed visiting scheme

The basic step towards determination of the mixing time on product spaces is the following

Proposition 2. *Let $k \geq 1$ and ρ be fixed. For transition matrices P_ρ we have*

$$d(P_\rho^k, Q_\rho^k) \leq \sum_{j=1}^d \mathrm{e}^{-\frac{k\rho_j}{8}} + \sum_{j=1}^d d_{\lfloor \frac{k\rho_j}{2} \rfloor + 1}(P_j).$$

Proof. Arguing as in the proof of Proposition 1 we obtain for any initial distribution μ on X a representation

$$\mu P_\rho^k - \mu Q_\rho^k = \sum_{r_1 + \dots + r_d = k} \binom{k}{\bar{r}} \mu \left(\prod_{j=1}^d \left(\rho_j \tilde{P}_j \right)^{r_j} - \prod_{j=1}^d \left(\rho_j \tilde{Q}_j \right)^{r_j} \right).$$

Taking into account that the transition matrices fulfill the properties from Lemma 4 we infer

$$\mu \prod_{j=1}^d \left(\rho_j \tilde{P}_j \right)^{r_j} - \mu \prod_{j=1}^d \left(\rho_j \tilde{Q}_j \right)^{r_j} =$$

$$\prod_{j=1}^d \rho_j^{r_j} \sum_{l=1}^d \left(\mu \prod_{j=1}^{l-1} \tilde{P}_j^{r_j} \right) \left(\tilde{P}_l - \tilde{Q}_l \right) \left(\prod_{j=l+1}^d \tilde{Q}_j^{r_j} \right)$$

(with obvious modification for $l = 1$ and $l = d$), which implies

$$\begin{aligned} d(P_\rho^k, Q_\rho^k) &\leq \sum_{r_1+\ldots+r_d=k} \binom{k}{\bar{r}} \prod_{j=1}^{d} \rho_j^{r_j} \sum_{l=1}^{d} d(\tilde{P}_l^{r_l}, \tilde{Q}_l^{r_l}) \\ &= \sum_{l=1}^{d} \sum_{r_l=0}^{k} \binom{k}{r_l} \rho_l^{r_l} (1-\rho_l)^{k-r_l} d(\tilde{P}_l^{r_l}, \tilde{Q}_l^{r_l}) \\ &\leq \sum_{l=1}^{d} \left(B_{k,\rho_l}(\left\{0,\ldots,\lfloor \frac{k\rho_l}{2} \rfloor\right\}) + \max_{r_l > \lfloor \frac{k\rho_l}{2} \rfloor} d_{r_l}(P_l) \right) \\ &\leq \sum_{l=1}^{d} \mathrm{e}^{-\frac{k\rho_l}{8}} + \sum_{l=1}^{d} d_{\lfloor \frac{k\rho_l}{2} \rfloor + 1}(P_l). \qquad (11) \end{aligned}$$

To derive the first sum in (11) we used the well known estimate

$$B_{k,p}(\left\{0,\ldots,\lfloor \frac{kp}{2} \rfloor\right\}) \leq \mathrm{e}^{-\frac{kp}{8}},$$

which is a consequence of Okamoto's result, see [4, Ch. 3.8]. The proof is complete. □

To proceed recall the definition of the mixing times $K(P)$ in (3). The main result in this section is

Theorem 1. *For any convex combination ρ we have*

$$K(P_\rho) \leq 8 \left(\max_{j=1,\ldots,d} \frac{K(P_j)}{\rho_j} \right) (1 + \lfloor 1 + \log(d) \rfloor) . \qquad (12)$$

Proof. Let $k \geq 8 \left(\max_{j=1,\ldots,d} \frac{K(P_j)}{\rho_j} \right) (1 + \lfloor 1 + \log(8d) \rfloor)$ be fixed. We have, using the results from Proposition 2 and Lemma 6, the following estimate.

$$\begin{aligned} d_k(P_\rho) &\leq d_k(Q_\rho) + d(P_\rho^k, Q_\rho^k) \\ &\leq 2 \sum_{l=1}^{d} \mathrm{e}^{-k\rho_l} + \sum_{l=1}^{d} \mathrm{e}^{-\frac{k\rho_l}{8}} + \sum_{l=1}^{d} d_{\lfloor \frac{k\rho_l}{2} \rfloor + 1}(P_l). \qquad (13) \end{aligned}$$

By our assumption on k the first and second sums above can be bounded by $\frac{1}{8\mathrm{e}}$. It can further be deduced from this assumption that $\frac{k\rho_l}{2} \geq (1 + \lfloor 1 + \log(8d) \rfloor) K(P_l)$, such that an application of Lemma 2 yields

$$d_{\lfloor \frac{k\rho_l}{2} \rfloor + 1}(P_l) \leq \frac{1}{8d\mathrm{e}}$$

from which the proof can be completed. □

4. Optimizing the visiting scheme

Below we allow to design our Markov chain P_ρ to fit the mixing properties of the components by varying ρ. This section is a straight–forward extension of the arguments provided in [5, Sect. 5].

As there we introduce the following notation. Given spaces X_j with Markov chains P_j having mixing times $K(P_j)$ we let

$$\kappa := \sum_{j=1}^{d} K(P_j) \quad \text{and} \quad \sigma_j := \frac{K(P_j)}{\kappa}, \quad j = 1, \dots, d.$$

The d–tuple $\sigma = (\sigma_1, \dots, \sigma_d)$ gives rise to a probability and we let

$$H(\sigma) := -\sum_{j=1}^{d} \sigma_j \log(\sigma_j)$$

denote the entropy of σ.

As for the setup in [5, Sect. 5] we have within the present context

Theorem 2. *Let*

$$\rho_j := \frac{\sigma_j(3 - \log(\sigma_j))}{H(\sigma) + 3},$$

such that this provides a convex combination $\bar{\rho}$. This specific combination $\bar{\rho}$ leads to

$$\inf_\rho K(P_\rho) \le K(P_{\bar{\rho}}) \le \lfloor 8\kappa(H(\sigma) + 3) \rfloor + 1. \tag{14}$$

The proof is the same as in [5, Sect. 5]. Of course, the above result lacks of an appropriate lower bound. As Lemma 6 suggests, some assumption on the richness of the components has to be made.

The bound (14) of the above Theorem is in fact a strengthening of Theorem 1, since $H(\sigma) \le \log(d)$ as well as $\kappa \le \max_{j=1,\dots,d} \frac{K(P_j)}{\rho_j}$.It is however surprising that the intuitively good choice of $\rho_j = \sigma_j, \quad j = 1, \dots, d$ does not lead to the same conclusion.

5. Application: Metropolis sampling with separable energy function

Within the framework studied above we apply the previous estimates to a Metropolis–type sampler. Thus we study the relaxation of product–type Markov chains which possess as invariant distributions a *Boltzmann distribution* Π_f. Such a distribution is defined through an (energy) function, say $f\colon X \to \mathbb{R}$ by letting

$$\Pi_f(\{x\}) := \frac{\mathrm{e}^{-f(x)}}{\sum_{y \in X} \mathrm{e}^{-f(y)}}, \quad x \in X.$$

Approximate sampling from Boltzmann distribution is the basis of Simulated Annealing, cf. [7]; such Markov chains which rapidly converge to the Boltzmann distribution allow global minimization without visiting most of the state space.

Of course, not every Boltzmann distribution can be approximated by product–type Markov chains, which points at serious limitations of the present approach.

However, if it can be approximated, then relaxation is achieved typically without visiting many states.

Metropolis–type Markov chains to approximately simulate the Boltzmann distribution are determined by an underlying Markov chain $\bar{P}$. Hence the compound transition matrix of the Metropolis Markov chain for an energy function f on a space X is

$$P_f(x,y) := \begin{cases} \mathrm{e}^{-(f(y)-f(x))^{+3}}\bar{P}(x,y) & \text{, if } y \neq x \\ 1-\sum_{z\in X\setminus\{x\}} \mathrm{e}^{-(f(z)-f(x))^{+}}\bar{P}(x,x) & , y = x \end{cases}.$$

Observe that $P_f(x,x) \geq \bar{P}(x,x)$ for obvious reasons. Moreover it is important that the invariant distribution of this Markov chain is the Boltzmann distribution Π_f, cf. [7, Ch. 8.2]. We shall concentrate on specific types of energies. Again we assume that the state space X is a product $X := \prod_{j=1}^d X_j$.

Definition 1. A function $f\colon \prod_{j=1}^d X_j \to \mathbb{R}$ shall be called *separable* if there exist functions $f_1,\dots,f_d$ acting on the components only such that

$$f(\xi_1,\dots,\xi_d) = \sum_{j=1}^d f_j(\xi_j), \quad (\xi_1,\dots,\xi_d) \in X.$$

The following is readily checked:

- The compound Boltzmann distribution Π_f is the product $\Pi_f = \prod_{j=1}^d \Pi_{f_j}$.
- If in addition the neighborhood system, which means the underlying Markov chain is of product type $\bar{P} = \frac{1}{d}\sum_{j=1}^d \tilde{P}_j$, then this is valid also for the compound Metropolis sampler P_f, i. e.,

$$P_f = \frac{1}{d}\sum_{j=1}^d \tilde{P}_{f_j}.$$

Within this context an application of Theorem 1 yields a constant such that the mixing time $K(P_f)$ can be estimated by

$$K(P_f) \leq C\left(\max_{j=1,\dots,d} K(P_{f_j})\right) d\log(d), \tag{15}$$

i.e., through the mixing times of the corresponding component Metropolis samplers, based on underlying Markov chains P_j, which remain to be estimated. This may be done under an additional assumption.

Definition 2. A Markov chain P on a space X is said to satisfy a *minorization condition*, if there is $\varepsilon > 0$ for which

$$\min_{\xi,\eta\in X} P(\xi,\eta) \geq \frac{\varepsilon}{|X|}. \tag{16}$$

Such condition is a powerful tool when studying convergence of Markov chains, we refer to [6, Sect. 6.2] for further details and references. The relevant result is

[3] for a real number r we let $r^+ := \max\{r, 0\}$.

Proposition 3. *The mixing time of the Metropolis–type sampler P_f based on a Markov chain satisfying a minorization for some ε can be bounded by*

$$K(P_f) \leq \frac{2\mathrm{e}^{\Delta_{max}}}{\varepsilon},$$

where $\Delta_{max} := \max_{\xi \in X} f(\xi) - \min_{\xi \in X} f(\xi)$.

Proof. The proof is based on estimating the contraction coefficient $\rho_1(P_f)$, see (4). In view of (5) we obtain

$$d_l(P_f) \leq \rho_l(P_f) \leq (\rho_1(P_f))^l,$$

such that it suffices to determine l for which $(\rho_1(P_f))^l \leq \frac{1}{2\mathrm{e}}$. The well–know estimate, see e. g. [7, Lemma 4.2.3], yields

$$\rho_1(P_f) \leq 1 - |X| \min_{\xi,\eta \in X} P_f(\xi, \eta).$$

This yields

$$\rho_1(P_f) \leq 1 - \frac{\mathrm{e}^{-\Delta_{max}}}{\varepsilon} \leq \mathrm{e}^{-\frac{\mathrm{e}^{-\Delta_{max}}}{\varepsilon}}.$$

The choice of $l \geq \frac{2\mathrm{e}^{\Delta_{max}}}{\varepsilon}$ finally provides $\rho_1(P_f)^l \leq \frac{1}{2\mathrm{e}}$. □

Applying this estimate to the component Metropolis samplers P_{f_j}, which are now supposed to be driven by Markov chains satisfying an ε–minorization, together with estimate (15) we obtain

Proposition 4. *If $f\colon \prod_{j=1}^d X_j \to \mathbb{R}$ is separable and the Metropolis sampler is based on a Markov chain of product type with components satisfying an ε–minorization condition, then there is a constant $0 < C < \infty$ for which*

$$K(P_f) \leq C \frac{\mathrm{e}^{\Delta_f}}{\varepsilon} d \log(d), \tag{17}$$

with $\Delta_f := \max_j(\max_{\xi \in \mathbb{Z}_n} f_j(\xi) - \min_{\xi \in \mathbb{Z}_n} f_j(\xi))$ being the maximal directional amplitude.

It is worth noting that this estimate is independent of the cardinality of the state space due to the minorization assumption. Best behavior from this point of view is predicted by sampling directly from the uniform distribution on each X_j, yielding $\varepsilon = 1$. This may contrast to the necessity of having a local underlying chain for fast computation of the differences $f_j(\eta) - f_j(\xi), \quad \xi, \eta \in X_j$.

One way to construct Markov chains satisfying a minorization is to chose a local random walk and let this relax until an appropriate minorization is achieved. The resulting compound Markov chain will then serve as underlying Markov chain for the Metropolis sampler. We are concerned with the problem, how long this relaxation takes. This can be solved using results from [2].

Our subsequent analysis requires additional notation, which is again close to the one from [2]. In addition to $d(P, \mu)$ as introduced in (2) we need the *separation distance* of a transition function P on a state space X[4] to its invariant distribution

[4] Again the space X is assumed to be arbitrary finite at this stage.

π by

$$s(P,\pi) := \max_{x,y\in X} |1 - \frac{P(x,y)}{\pi(\{y\})}|. \tag{18}$$

For $\delta > 0$ we let

$$S_\delta(P) := \min\left\{k, \quad s(P^k,\pi) \le \delta\right\} \tag{19}$$

be the minimal number of transitions of P required to make the distribution at the k–th step δ–close to the invariant distribution π.

Recall that $K(P)$ denotes the mixing time and that the invariant distribution for a symmetric transition function is necessarily the uniform one. In view of [2, Prop. 5. 13] we have

Lemma 7. *Let P be a symmetric transition function with mixing time $K(P)$. Then we have*

$$S_\delta(P) \le 2K(P)(1 + \lfloor\log(\frac{32}{\delta^2})\rfloor). \tag{20}$$

Proof. Let $k_\delta = K(P)(1 + \lfloor\log(\frac{32}{\delta^2})\rfloor)$. In view of Lemma 2 we have

$$d_{k_\delta}(P) \le \frac{\delta^2}{32}.$$

By Proposition 5.13 from [2] we can bound

$$s(P^k,U) \le 4\sqrt{2d_{k_\delta}(P)} \le \delta,$$

whenever $k \ge 2k_\delta$, completing the proof. □

This leads to

Corollary 1. *Let P be a symmetric transition function on a space X with mixing time $K(P)$ towards the invariant distribution U. For $k \ge 2K(P)(1 + \lfloor\log(\frac{32}{(1-\varepsilon)^2})\rfloor)$ we have*

$$\min_{\xi,\eta\in X} P^k(\xi,\eta) \ge \frac{\varepsilon}{|X|}.$$

Proof. Under the assumption on k we apply Lemma 7 to bound

$$s(P^k,U) \le 1 - \varepsilon$$

and an application of the triangle inequality yields finally for arbitrary ξ and η in X

$$P^k(\xi,\eta) \ge \varepsilon \min_{\xi\in X} U(\{\xi\}) \ge \frac{\varepsilon}{|X|}.$$

□

Returning to the original setup of Metropolis samplers for separable energy functions on product spaces we state that for letting each component Markov chain be $P_j := P^{10K(P)}$, such that one P_j step is $10K(P)$ steps according to the nearest neighbor walk P, then each P_j obeys a minorization condition with $\varepsilon \geq \frac{1}{5}$.

Summarizing, let us briefly discuss a Metropolis sampler on a grid $\mathbb{Z}_n^d$ for a separable energy function based on component nearest neighbor walks. The mixing time of such nearest neighbor walk is known to be proportional to $n^2/2$, see [3, Ch. 3C]. The above analysis yields that $50\mathrm{e}^{\Delta_f} n^2$ steps suffice for the component Markov chains to approach stationarity. An application of estimate (15) implies a constant C for which $C\mathrm{e}^{\Delta_f} n^2 d \log(d)$ steps suffice to approach stationarity of P_f. In conclusion, the portion r of states visited to the overall number n^d of states is bounded by

$$r \leq \frac{C\mathrm{e}^{\Delta_f} n^2 d \log(d)}{n^d},$$

which is small for at least moderate values of d and n, provided Δ_f was not too large, say $\Delta_f << d\log(n)$.

Hence for separable functions the Boltzmann distribution can be approximated using the Metropolis sampler on product spaces without visiting most states especially in high dimensions.

References

[1] D. Aldous. Random walks on finite groups and rapidly mixing markov chains. In *Seminaire de Probabilites XVII*, volume 986 of *Lect. Notes Math.*, pages 243 – 297, Berlin, Heidelberg, New York, London, Paris, Tokyo, 1983. Springer.

[2] D. Aldous and P. Diaconis. Strong uniform times and finite random walks. *Adv. in Appl. Math.*, 8:69 – 97, 1987.

[3] P. Diaconis. *Group Representations in Probability and Statistics*, volume 11 of *Lect. Notes–Monogr. Series*. Instit. Math. Statist., Hayward, 1988.

[4] N. L. Johnson and S. Kotz. *Distributions in Statistics: Discrete Distributions*. John Wiley & Sons, New York, Chichester, Brisbane, 1969.

[5] P. Mathé. Efficient mixing of product walks on product groups. preprint WIAS, December 1996.

[6] J. S. Rosenthal. Convergence rates for Markov chains. *SIAM Rev.*, 37(3):387–405, 1995.

[7] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, volume 27 of *Appl. Math.* Springer, Berlin, Heidelberg, 1995.

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D–10117 Berlin, Germany

E-mail address: mathe@wias-berlin.de