An Elementary Rigorous Introduction to Exact Sampling

F. Friedrich G. Winkler O. Wittich
V. Liebscher

Institute of Biomathematics and Biometry
GSF - National Research Center for Environment and Health,
Postfach 1129, D-85758 Oberschleißheim, Germany,
{friedrich,gwinkler,wittich,liebscher}@gsf.de,
http://www.gsf.de/institute/ibb/

Abstract

We introduce coupling from the past, a recently developed method for exact sampling from a given distribution. Focus is on rigour and thorough proofs. We stay on an elementary level which requires little or no prior knowledge from probability theory. This should fill an obvious gap between innumerable intuitive and incomplete reviews, and few precise derivations on an abstract level.

1 Introduction

We introduce a recently developed method for exact sampling from a given distribution. It is called *coupling from the past*. This is in contrast to Markov chain Monte Carlo samplers like the Gibbs sampler or the family of Metropolis-Hastings samplers which return samples from a distribution approximating the target distribution. The drawback is that MCMC methods apply generally and exact sampling works in special cases only. On the other hand, it is the object of current research and the list of possible applications increases rapidly. Another advantage is that problems like burn in and convergence diagnostics do not arise where exact sampling works. Exact sampling was proposed in the seminal paper J.G. Propp and D.B. Wilson (1996). Whereas these authors called the method *exact sampling*, some prefer the term *perfect sampling* since random sampling never is exact. For background in Markov chains and sampling, and for examples, we refer to G. Winkler (1995, 2003).

The aim of the present paper is a rigorous derivation and a thorough analysis at an elementary level. Nothing is really new; the paper consists of a combination of ideas, examples, and techniques from various recent papers, basically along the lines in F. Friedrich (2003). Hopefully, we can single out the basic conditions under which the method works theoretically, and what has to be added for a practicable implementation.

Coupling from the past is closely related to Markov Chain Monte Carlo sampling (MCMC), which nowadays is a widespread and commonly accepted statistical tool, especially in Bayesian statistical analysis. Hence we premise the discussion of coupling to the past with some remarks on Markov Chain Monte Carlo sampling. Let us first introduce the general framework which simultaneously gives us the basis for coupling from the past. For background and a detailed discussion see G. Winkler (1995), [13].

Let $\boldsymbol{X}$ be a finite set of generic elements $x, y, \ldots$. A *probability distribution* ν on $\boldsymbol{X}$ is a function on $\boldsymbol{X}$ taking values in the unit interval $[0,1]$ such that $\sum_{x\in\boldsymbol{X}}\nu(x) = 1$. A *Markov kernel* or *transition probability* on $\boldsymbol{X}$ is a function $P : \boldsymbol{X}\times\boldsymbol{X} \to [0,1]$ such that for each $x \in \boldsymbol{X}$ the function $P(x,\cdot) : \boldsymbol{X} \to [0,1]$, $y \mapsto P(x,y)$ is a probability distribution on $\boldsymbol{X}$. A probability distribution ν on $\boldsymbol{X}$ can be interpreted as a row vector $(\nu(x))_{x\in\boldsymbol{X}}$ and a Markov kernel P as a stochastic matrix $(P(x,y))_{x,y\in\boldsymbol{X}}$. A *right Markov chain* with initial distribution ν and transition probability P is a sequence $(\xi_i)_{i\geq 0}$ of random variables the law of which is determined by ν and P via the finite-dimensional marginal distributions given by

$$\mathbb{P}\big(\xi_0 = x_0, \xi_1 = x_1, \ldots, \xi_n = x_n\big) = \nu\big(x_0\big)P\big(x_0, x_1\big)\cdot\cdots\cdot P\big(x_{n-1}, x_n\big).$$

P is called *primitive* if there is a natural number τ such that $P^\tau(x,y) > 0$ for all $x, y \in \boldsymbol{X}$. This means that the τ-step probability from state x to state y is strictly positive for arbitrary x and y. If P is primitive then there is a unique probability distribution μ which is *invariant* w.r.t. P, i.e. $\mu P = \mu$ where μP is the matrix product of the (left) row vector μ and the matrix P, and this invariant probability distribution μ is strictly positive.

The laws or distributions of the variables ξ_n of such a process converge to the invariant distribution, i.e.

$$\nu P\cdot\cdots\cdot P(y) \longrightarrow \mu(y),\;\; y \in \boldsymbol{X}, \tag{1}$$

cf. [13], Theorem 4.3.1. Perhaps the most important statistical features to be estimated are expectation values of functions on the state space $\boldsymbol{X}$, and the most common estimators are empirical means. Fortunately, such stochastic processes fulfill the law of large numbers, which in its most elementary version reads: For each function f on $\boldsymbol{X}$, the empirical means along time converge in probability (and in L^2) to the expectation of f with respect to the invariant distribution; in formulae this reads

$$\frac{1}{n}\sum_{i=0}^{n-1} f(\xi_i) \longrightarrow \mathbb{E}(f;\mu) \;\text{ as } n\to\infty, \quad \text{in probability}, \tag{2}$$

(cf. [13], Theorem 4.3.2). The symbol $\mathbb{E}(f;\mu)$ denotes the expectation

$$\mathbb{E}\big(f;\mu\big) = \sum_{x\in\boldsymbol{X}} f(x)\mu(x)$$

of f with respect to μ. A sequence of random variables ξ_i *converges* to the random variable ξ *in probability* if for each $\varepsilon > 0$ the probability $\mathbb{P}(|\xi_i - \xi| > \varepsilon)$ tends to 0 as n tends to ∞. Plainly, (2) implies that for every natural number m, averaging may be started from m without destroying convergence in probability; more precisely for each $m \geq 0$ one has

$$\frac{1}{n-m} \sum_{i=m+1}^{n} f(\xi_i) \longrightarrow \mathbb{E}(f;\mu) \text{ as } n \to \infty, \quad \text{in probability.} \tag{3}$$

In view of the law of large numbers for identically distributed and independent variables, the step number m should be large enough such that the distributions of the variables $\xi_{m+1}, \ldots, \xi_n$ are close to the invariant distribution μ in order to estimate the expectation of f with respect to μ properly from the samples $f(\xi_{m+1}), \ldots, f(\xi_n)$.

In fact, according to (1), after some time m the laws of the ξ_i should be close to the invariant distribution μ although they may be far from μ during the initial period. The values during this *burn in* period are usually discarded and an average $(\sum_{m+1}^{n} f(\xi_i))/(n-m)$ like in (3) is computed. In general, the burn in time can hardly be determined. There are a lot of suggestions ranging from visual inspection of the time series $(f(\xi_i))_{i\geq 0}$ to more formal tools, called *convergence diagnostics*. In this text we are not concerned with burn in and restrict ourselves to the illustration in Fig. 1. A Gibbs sampler (introduced in Section 4) for the Ising model is started with a pepper and salt configuration in the left picture. A typical sample of the invariant distribution is the right one which appears after about 8000 steps. The pictures in-between show intermediate configurations which are pretty improbable given the invariant distribution but which are quite stable with respect to the Gibbs sampler. In physical terms, the right middle configuration is close to a 'meta-stable' state. Since we are interested in a typical configuration of the invariant distribution μ, we should consider the burn in to be completed if the sample from the Markov chain looks like the right hand side of Fig. 1, i.e. after about 8000 steps of the Gibbs sampler. The curve in the next figure Fig. 2 displays the relative frequency of

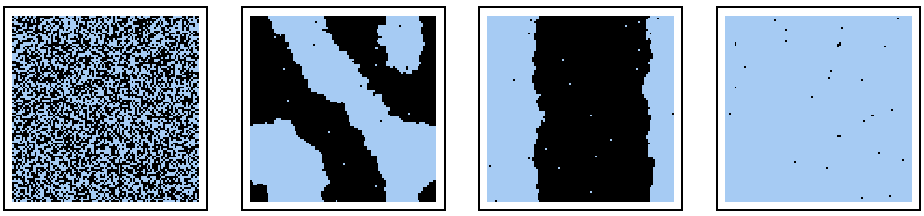


Figure 1: Configurations for Ising Gibbs Sampler with $\beta = 0.8$ starting in a pepper and salt-configuration (left), after 150 steps (left middle), after 350 steps (right middle) and after 8000 steps (right).

equal neighbour pairs. Superficial visual inspection of this plot suggests that the sampler should be in equilibrium after about 300 steps. On the other hand, comparison with Fig. 1 reveals that the slight ascent at about 7800 steps presumably is much more relevant for the decision whether burn is completed or not. This indicates that primitive diagnostic tools may be misleading. The interested reader is referred to the references in [7; 6; 10], see W.R. Gilks et al. (1996b). If initial samples from μ itself are available, then there is no need for

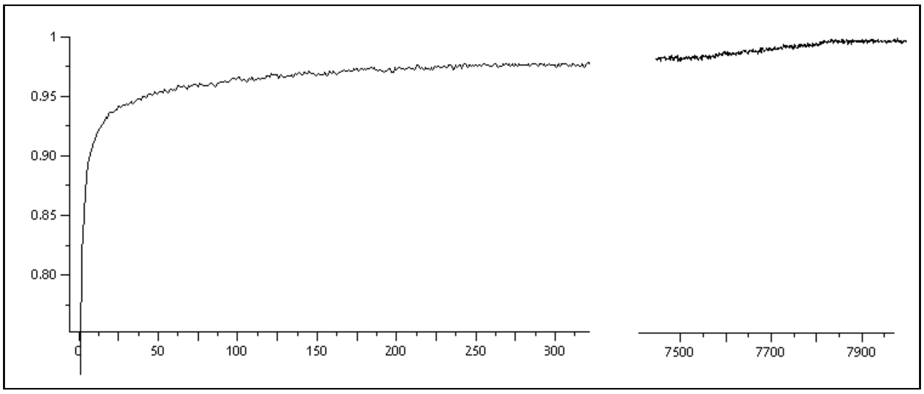
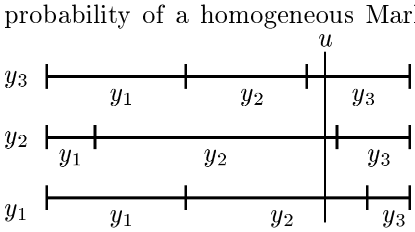


Figure 2: Convergence Diagnostics for Ising Gibbs Sampler

a burn in, and one can average from the beginning. This is one of the most valuable advantages of exact sampling.

First, we indicate how a Markov chain can be simulated.

Example 1 (Simulating a Markov chain) We denote by P the transition probability of a homogeneous Markov chain. At each time $n \geq 1$, given the previous state x_{n-1}, we want to pick a state x_n at random from $P(x_{n-1}, \cdot)$. For each x, we partition the unit interval $(0,1]$ into intervals I_y^x of length $P(x,y)$, and pick u_n uniformly at random from $(0,1]$. Given the present state x_{n-1}, we search for the state y with $u_n \in I_y^{x_{n-1}}$ and set $x_n = y$. The picture on the left illustrates this procedure for $|\boldsymbol{X}| = 3$, where $x_n = y_2$ if x_{n-1} was y_1 or y_2 and $x_n = y_3$ if $x_{n-1} = y_3$. In general, the procedure can be rephrased as follows: Define a *transition rule* for P by



$$f : \boldsymbol{X} \times (0,1] \longrightarrow \boldsymbol{X}, \;\; f(x,u) = y \quad \text{if and only if} \quad u \in I_y^x.$$

More explicitly, enumerate $\boldsymbol{X} = \{y_1, \ldots, y_N\}$ and set $f(x,u) = F_x^-(u)$ where $F_x(u) = P(x, \{y_i : i \leq u\})$ is the cumulative distribution function of $P(x,\cdot)$ and $F_x^-(u) = \min\{t : F_x(t) \geq u\}$ its generalized inverse. Let $U_1, U_2, \ldots$ be independent random variables uniformly distributed over $(0,1]$, and set $\xi_0 := x_0$, and $\xi_n := f(\xi_{n-1}, U_n)$. Then $(\xi_n)_{n\geq 0}$ is a homogeneous Markov chain starting at x_0 with transition probability P. For inhomogeneous chains, replace

f by f_n varying in time. Note that the exclusive source of randomness are the independent random variables U_i.

2 Exact Sampling

The basic idea of coupling from the past is closely related to the law of large numbers (2). According to (1), for primitive P with invariant distribution μ the corresponding Markov chain converges to μ; more precisely

$$\nu P^n \longrightarrow \mu, \text{ as } n \to \infty, \tag{4}$$

uniformly in all initial distributions ν, and with respect to any norm on $\mathbb{R}^{\boldsymbol{X}}$.

Generalizing the concept of right Markov chains, let us consider now *two-sided Markov chains* with transition probabilities given by a Markov kernel P, i.e. double sequences $(\xi_i)_{i\in\mathbb{Z}}$ of random variables taking values in $\boldsymbol{X}$, and with law determined by the marginal distributions

$$\mathbb{P}\big(\xi_m = x_m, \dots, \xi_n = x_n\big) = \nu_m\big(x_m\big)P\big(x_m, x_{m+1}\big) \cdot \dots \cdot P\big(x_{n-1}, x_n\big), \tag{5}$$

for $m, n \in \mathbb{Z}$, $n > m$, where ν_k denotes the law of ξ_k.

If P is primitive, or more generally, if (4) holds uniformly, these two-sided chains are automatically *stationary*. This important concept means that a time shift does not change the law of the chain; in terms of the marginal distributions this reads

$$\mathbb{P}\big(\xi_m = x_m, \dots, \xi_n = x_n\big) = \mathbb{P}\big(\xi_{m+\tau} = x_m, \dots, \xi_{n+\tau} = x_n\big) \tag{6}$$

for all $m \in \mathbb{Z}$ and $\tau \in \mathbb{Z}$, and in particular, that all ν_m in (5) are equal to μ. In fact, because of (5) one has $\nu_0 = \nu_{-k}P^k$ for all $k \in \mathbb{N}$. By uniformity in (4), this implies $\nu_0 = \mu$ and hence in view of (5) the process $(\xi_i)_{i\in\mathbb{Z}}$ is stationary.

At a first glance, this does not seem to be helpful since we cannot simulate the two-sided chain starting at time $-\infty$. On the other hand, if we want to start sampling at some (large negative) time n, there is no distinguished state to start in, since stationarity of the chain implies that the initial state necessarily is already distributed according to μ. The main idea to overcome this problem is to start chains simultaneously at all states in $\boldsymbol{X}$ and at each time. This means that a lot of Markov chains are *coupled* together. The coupling will be constructed in such a fashion that if two of the chains happen to be in the same state in $\boldsymbol{X}$ at some (random) time, they will afterwards follow the same trajectory forever. This phenomenon is called *coalescence* of trajectories. Our definite aim is to couple the chains in a cooperative way such that after a large time it is very likely that any two of the chains have met each other at time 0. Then, at time 0, all chains started simultaneously at sufficiently large negative time have coalesced, and therefore their common state at time 0 does not dependent on the starting points in the far past anymore. We will show that after complete coalescence the unique random state at time 0 *is distributed according to the invariant distribution* μ.

To make this precise we consider the following setup: Let $\boldsymbol{X}$ be a finite space and let μ be a strictly positive probability distribution on $\boldsymbol{X}$. The aim is to realize a random variable which *exactly* has law μ, or - in other words - to sample from μ. Since Markov chains have to be started at each time $k<0$ and at each state $x \in \boldsymbol{X}$ simultaneously, a formal framework is needed into which all these processes can be embedded. The appropriate concept is that of *iterated random maps* or *stochastic flows*, systematically exploited in P. DIACONIS AND D. FREEDMAN (1999).

Let μ be the strictly positive distribution on $\boldsymbol{X}$ from which we want to sample and let P be a Markov kernel on $\boldsymbol{X}$ for which μ is the unique invariant distribution. Let Φ be the set of all maps from $\boldsymbol{X}$ to itself:

$$\Phi = \{\varphi : \boldsymbol{X} \longrightarrow \boldsymbol{X}\} = \boldsymbol{X}^{\boldsymbol{X}} = \mathrm{Map}(\boldsymbol{X}, \boldsymbol{X}).$$

On this space we consider distributions p reflecting the action of P on $\boldsymbol{X}$ in the sense that the p-probability that some point x is mapped by the random function φ to some y is given by $P(x,y)$. This connection between p and P is formalized by the condition

$$\text{(P)} \qquad p\left(\{\varphi : \varphi(x) = y\}\right) = P(x,y),\ x,y \in \boldsymbol{X}.$$

Example 2 Such a distribution does always exist. A synchronous one is given by $q(\varphi) = \prod_{x \in \boldsymbol{X}} P(x, \varphi(x))$. It is a probability distribution since it can be written as a product of the distributions $P(x,\cdot)$. It also fulfills Condition (P): Let Φ' be the set of all maps from $\boldsymbol{X}\backslash\{x\}$ to $\boldsymbol{X}$. Then

$$\begin{aligned} & q\big(\varphi : \varphi(x) = y\big) \\ & \qquad = \sum_{\{\varphi : \varphi(x) = y\}} \prod_{z \in \boldsymbol{X}} P(z, \varphi(z)) = P(x,y) \sum_{\varphi \in \Phi'} \prod_{z \neq x} P(z, \varphi(z)) = P(x,y); \end{aligned}$$

the sum over Φ' equals 1 since the summands again define a product measure.

Since we want to mimic Markov processes, we need measures on sets of paths, and since we will proceed from time $-\infty$ to finite times we introduce measures on the set $\Omega = \Phi^{\mathbb{Z}}$ with one-dimensional marginal measures p. The simplest choice are product measures $\mathbb{P} = p^{\mathbb{Z}}$. The space $\Omega = \Phi^{\mathbb{Z}}$ consists of double sequences

$$\underline{\varphi} = (\varphi_j)_{j \in \mathbb{Z}} = (\ldots, \varphi_{-1}, \varphi_0, \varphi_1, \ldots) \in \mathrm{Map}(\boldsymbol{X}, \boldsymbol{X})^{\mathbb{Z}}.$$

If J is a finite subset of $\mathbb{Z}$ then for each choice ψ_j, $j \in J$, we have

$$\mathbb{P}\big(\{\underline{\varphi} \in \Omega : \varphi_j = \psi_j,\ j \in J\}\big) = \prod_{j \in J} p(\psi_j).$$

Given a double sequence $\underline{\varphi}$ of maps φ_j, $j \in \mathbb{Z}$, we consider compositions of the components φ_j over time intervals. For each $\underline{\varphi} \in \Omega$ and $x \in \boldsymbol{X}$, set

$$\varphi_j^k(x) = \varphi_k \circ \cdots \circ \varphi_j(x) = \varphi_k\big(\varphi_{k-1}\big(\cdots(\varphi_j(x))\big)\big),\ j \leq k.$$

Note that $\varphi_i^i = \varphi_i$.

Remark Given Condition (P), for each $n \in \mathbb{Z}$ and $x \in \boldsymbol{X}$, the process $\xi_n \equiv x$, $\xi_{n+k} = \varphi_{n+1}^{n+k}(x)$, $k \geq 1$, is a Markov chain starting at x and with transition probability P. Hence the stochastic flow is a common representation of Markov chains starting at all initial states and at all times; we shall say that they are *coupled from the past.*

Coupling from the past at time n will work as follows: Pick a double sequence

$$\dots, \varphi_m, \dots, \varphi_n, \dots$$

of maps at random, and fix a number $n \in \mathbb{Z}$. Then decrease m until $\varphi_m^n(x) = w$ hopefully does not depend on x anymore. If we are successful and this happens then we say that all trajectories

$$\varphi_m(x), \varphi_{m+1} \circ \varphi_m(x), \dots, \varphi_m^n(x), \;\; x \in \boldsymbol{X},$$

have *coalesced.* We shall also say that for $\underline{\varphi}$ there is *complete coalescence* at time n. This works if sufficiently many of the φ_j map different elements x to the same image. Going further backwards does not change anything since $\varphi_{m-k}^n(x) = \varphi_m^n(\varphi_{m-k}^{m-1}(x)) = w$ holds as well for all x. This may be rephrased in terms of sets as follows: Let $\varphi : \boldsymbol{X} \to \boldsymbol{X}$ be a map and $\mathrm{Im}\varphi = \{\varphi(x) : x \in \boldsymbol{X}\}$ the image of $\boldsymbol{X}$ under φ. For fixed n the sets $\mathrm{Im}\varphi_m^n$ decrease as m decreases. Complete coalescence means that $\mathrm{Im}\varphi_m^n$ is a singleton $\{w\}$. Then there is a unique $W_n(\underline{\varphi}) \in \boldsymbol{X}$ with

$$\{W_n(\underline{\varphi})\} := \bigcap_{m \leq n} \mathrm{Im}\varphi_m^n. \tag{7}$$

If there is no coalescence then $W_n(\underline{\varphi})$ is not defined. Let us set

$$F_n = \{\underline{\varphi} : W_n(\underline{\varphi}) \text{ exists}\}, \;\; F = \bigcap_{n \in \mathbb{Z}} F_n.$$

Then all W_n are well defined on F; to complete the definition let $W_n(\underline{\varphi}) = z_0$ for some fixed $z_0 \in \boldsymbol{X}$ if $\underline{\varphi} \notin F$. Obviously, independent of the choice of $x \in \boldsymbol{X}$,

$$\begin{aligned} W_{n+k}(\underline{\varphi}) &= \lim_{m \to -\infty} \varphi_m^{n+k}(x) = \varphi_{n+1}^{n+k} \circ \lim_{m \to -\infty} \varphi_m^n(x) \\ &= \varphi_{n+1}^{n+k} \circ W_n(\underline{\varphi}), \qquad\qquad \underline{\varphi} \in F, \;\; n \in \mathbb{Z}, \;\; k > 0. \end{aligned} \tag{8}$$

This indicates that the random variables $W_n(\underline{\varphi})$ have law μ. To exploit this observation for a sampling algorithm we need almost sure complete coalescence in finite time. We enforce this by the formal condition

$$\text{(F)} \qquad \mathbb{P}(F) = 1.$$

Provided that (F) holds, we call $\mathbb{P}$ *successful.* Condition (F) will be verified below under natural conditions.

Lemma 1 *Under the hypothesis* (P) *and* (F) *the process* $(W_m)_{m\in\mathbb{Z}}$ *is a stationary homogeneous Markov process with Markov kernel* P.

Proof. Recall that $\mathbb{P}$ is a homogeneous product measure, and hence for each $\tau \in \mathbb{Z}$ all random sequences $\varphi_m, \ldots, \varphi_{m+\tau}$, $m \in \mathbb{Z}$, have the same law. Hence the stochastic flow is stationary, and the process $(W_m)_{m\in\mathbb{Z}}$ is stationary as well. Moreover, φ_{n+1}^{n+k} depends on $\varphi_{n+1}, \ldots, \varphi_{n+k}$ only and each W_m depends only on $\ldots, \varphi_{m-1}, \varphi_m$. Again, since $\mathbb{P} = p^{\mathbb{Z}}$ is a product measure, the variables φ_{n+1}^{n+k} and W_m, $m \le n$, are independent. By (8) and (P),

$$\begin{aligned}
&\mathbb{P}\big(W_{n+1} = x_{n+1}, W_n = x_n, \ldots, W_{n-k} = x_{n-k}\big)\\
&\quad = \mathbb{P}\big(\varphi_{n+1}^{n+1}(x_n) = x_{n+1}, W_n = x_n, \ldots, W_{n-k} = x_{n-k}\big)\\
&\quad = \mathbb{P}\big(\varphi_{n+1}(x_n) = x_{n+1}\big)\mathbb{P}\big(W_n = x_n, \ldots, W_{n-k} = x_{n-k}\big)\\
&\quad = P\big(x_n, x_{n+1}\big)\mathbb{P}\big(W_n = x_n, \ldots, W_{n-k} = x_{n-k}\big),
\end{aligned}$$

which shows

$$\mathbb{P}\big(W_{n+1} = x_{n+1} \big| W_n = x_n, \ldots, W_{n-k} = x_{n-k}\big) = P(x_n, x_{n+1}\big).$$

Hence P is the transition probability of the process $(W_m)_{m\in\mathbb{Z}}$. $\square$

Let us put things together in the first main theorem.

Theorem 1 (Exact Sampling) *Suppose that* μ *is a strictly positive probability distribution and* P *a primitive Markov kernel on* $\boldsymbol{X}$ *such that* $\mu P = \mu$. *Assume further that* $p\left(\{\varphi : \varphi(x) = y\}\right) = P(x,y)$ *for all* $x, y \in \boldsymbol{X}$, *and that* $\mathbb{P}$ *is successful. Then each random variable* W_n *has law* μ*; more precisely:*

$$\mathbb{P}\left(\{\underline{\varphi} \in \Omega : W_n(\underline{\varphi}) = x\}\right) = \mu(x),\ \ x \in \boldsymbol{X}. \tag{9}$$

Proof. By stationarity from Lemma 1, all one-dimensional marginal distributions coincide, and P is the transition probability of $(W_n)_{n\in\mathbb{Z}}$. If P is primitive then by [13], Theorem 4.3.1, its unique invariant distribution is μ. $\square$

To sample from μ, only one of the W_m is needed.

Corollary *Under the assumptions of Theorem 1, the random variable* W_0 *has law* μ.

The next natural question concerns the waiting time for complete coalescence at time zero. The random times T_n of latest coalescence before n are given by

$$T_n(\underline{\varphi}) = \sup\{m \le n : \text{there is } w \in \boldsymbol{X} \text{ such that } \varphi_m^n(x) = w \text{ for every } x \in \boldsymbol{X}\}.$$

The numbers $T_n(\underline{\varphi})$ definitely are finite if $\underline{\varphi} \in F$; outside F they may be finite or equal $-\infty$. Condition (F) is equivalent to

$$\mathbb{P}\big(\{\underline{\varphi} \in \Omega : T_n(\underline{\varphi}) > -\infty\}\big) = 1 \quad \text{for every} \quad n \in \mathbb{Z}. \tag{10}$$

Such a random time is also called *successful.* To realize W_0 one subsequently and independently picks maps $\varphi_0, \varphi_{-1}, \ldots, \varphi_m$ until there is coalescence say in $w \in \boldsymbol{X}$. This element w is a sample from μ. For computational reasons, one usually goes back in time by powers of 2. Clearly, choosing $k_0(\underline{\varphi})$ such that $-2^{k_0(\underline{\varphi})} \leq T_n(\underline{\varphi})$ assures coalescence at time 0. Recall that such a $k_0(\underline{\varphi})$ exists for each $\underline{\varphi} \in F$. An example of a stochastic flow coalescing completely at time $m = 0$ is shown in Fig. 3. We are going now to discuss a condition for (F) to

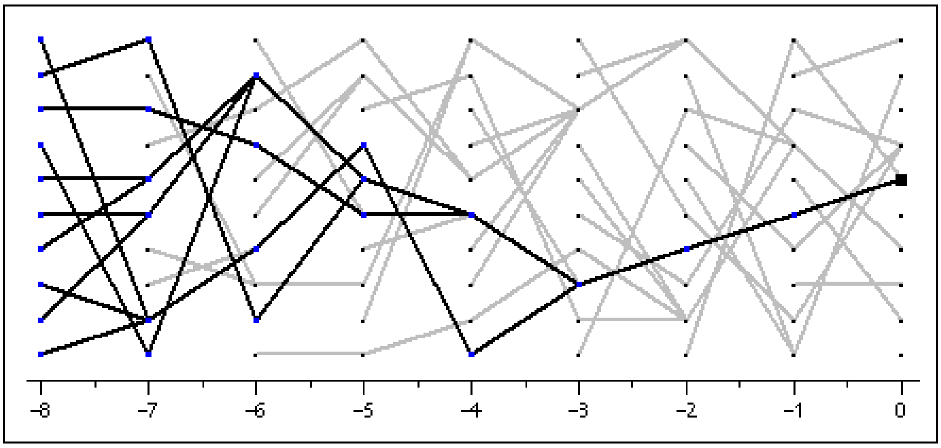


Figure 3: Latest complete coalescence time before time 0

hold. *Pairwise coalescence with positive probability* is perhaps the most natural condition and easy to check:

(C) For each pair $x, y \in \boldsymbol{X}$ there is an integer $n(x,y)$ such that

$$p^{\,n(x,y)}\big(\{(\varphi_1, \ldots, \varphi_{n(x,y)}) \in \Phi^{n(x,y)} : \varphi_1^{n(x,y)}(x) = \varphi_1^{n(x,y)}(y)\}\big) > 0.$$

We shall show in Theorem 2 below that (C) and (F) are equivalent. We give now a simple example where coupling fails.

Example 3 Consider P with invariant μ on $\boldsymbol{X} = \{1,2\}$ given by

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}, \quad \mu = \big(1/2, 1/2\big).$$

Let $p(\iota) = 1/2 = p(\psi)$ for the identity map $\iota(1) = 1$, $\iota(2) = 2$, and the flip map $\psi(1) = 2$, $\psi(2) = 1$. Compositions of ι and ψ never will couple. On the other hand the flow is associated to P since $p(\{\varphi : \varphi(x) = y\}) = 1/2 = P(x,y)$, regardless of x and y, and Condition (P) holds.

We shall show now that the coupling condition (C) implies complete coalescence (F) (and the converse). The latter condition may be rephrased as follows: All random times T_n are finite almost surely. By stationarity this boils down to: The random time T_0 is finite almost surely. The simplest, but fairly abstract way to verify (F) is to use shift invariance of F and ergodicity of $\mathbb{P}$. We will argue along these lines but in a more explicit and elementary way. The first step is to ensure existence of a finite τ such that the flow coalesces completely in less than τ steps with positive probability.

Lemma 2 *Under condition* (C) *there is a natural number τ such that*

$$\mathbb{P}\big(\{\underline{\varphi} : T_0(\underline{\varphi}) > -\tau\}\big) > 0.$$

Proof. Let $n_c = \max\{n(x,y) : x, y \in \boldsymbol{X}\}$. If $\varphi_1^n(x) = \varphi_1^n(y)$ for some $n < n_c$ then $\varphi_1^{n_c}(x) = \varphi_{n+1}^{n_c} \circ \varphi_1^n(x) = \varphi_1^{n_c}(y)$ as well. Hence Condition (C) implies

$$q = \min\Big\{p^{n_c}\big\{(\varphi_1, \ldots, \varphi_{n_c}) : \varphi_1^{n_c}(x) = \varphi_1^{n_c}(y)\big\} : x, y \in \boldsymbol{X}\Big\} > 0.$$

Therefore $|\boldsymbol{X}| > |\mathrm{Im}\varphi_1^{n_c}|$ at least with probability $q > 0$ if $|\boldsymbol{X}| \geq 2$. Similarly, $|\mathrm{Im}\varphi_1^{n_c}| > |\mathrm{Im}\varphi_1^{2n_c}|$ with probability at least q^2 if the left set is no singleton. This holds because $\varphi_1^{2n_c} = \varphi_{n_c+1}^{2n_c} \circ \varphi_1^{n_c}$ and the variables $\varphi_1, \ldots, \varphi_{n_c}$ and $\varphi_{n_c+1}, \ldots, \varphi_{2n_c}$ are independent and identically distributed. By induction,

$$\big|\boldsymbol{X}\big| > \big|\mathrm{Im}\varphi_1^{n_c}\big| > \big|\mathrm{Im}\varphi_1^{2n_c}\big| > \cdots > \big|\mathrm{Im}\varphi_1^{kn_c}\big|$$

at least with probability q^k until the last cardinality becomes 1; this happens after at most $|\boldsymbol{X}| - 1$ steps. Let $\tau = (|\boldsymbol{X}| - 1)n_c - 1$. Nothing changes if we renumber the maps as $\varphi_{-\tau}, \ldots, \varphi_0$, $m < 0$. Hence $\mathbb{P}(\{|\mathrm{Im}\varphi_{-\tau}^0| = 1\}) \geq q^\tau$ and the lemma is proved. $\square$

The next step is a sub-multiplicativity property of probabilities for coalescence times.

Lemma 3 *Let $n, m < 0$ be negative integers. Then*

$$\mathbb{P}\big(T_0 \leq m+n\big) \leq \mathbb{P}\big(T_0 \leq m\big)\mathbb{P}\big(T_m \leq m+n\big) = \mathbb{P}\big(T_0 \leq m\big)\mathbb{P}\big(T_0 \leq n\big).$$

Proof. Suppose that $T_0(\underline{\varphi}) \leq m+n$. This holds if and only if $\mathrm{Im}\varphi_{m+n+1}^0$ has more than one element. Then both, $\mathrm{Im}\varphi_{m+1}^0$ and $\mathrm{Im}\varphi_{m+n+1}^m$, have more than one element. Hence

$$\mathbb{P}\big(\underline{\varphi} : T_0(\underline{\varphi}) \leq m+n\big) \leq \mathbb{P}\big(\underline{\varphi} : T_0(\underline{\varphi}) \leq m \text{ and } T_m(\underline{\varphi}) \leq m+n\big).$$

To check whether $T_0(\underline{\varphi}) \leq m$ holds true it is sufficient to know the maps $\varphi_{m+1}, \ldots, \varphi_0$, and similarly, to check $T_m(\underline{\varphi}) \leq n+m$ only $\varphi_{m+n+1}, \ldots, \varphi_m$ are needed. Hence the respective sets are independent and the inequality holds. The remaining identity follows from stationarity. $\square$

In combination with Theorem 1, the next result completes the derivation of exact sampling.

Theorem 2 *The Conditions* (F) *and* (C) *are equivalent. In particular, the process governed by* $\mathbb{P}$ *is successful under* (C)*, and almost sure coalescence in Theorem 1 is assured.*

Proof. Suppose that (C) holds. By Lemma 2, we have $\mathbb{P}(T_0 > -\tau) > 0$ and Lemma 3 implies

$$\mathbb{P}(T_0 \le -n\tau) \le \mathbb{P}(T_0 \le -\tau)^n = \big(1 - \mathbb{P}(T_0 > -\tau)\big)^n \longrightarrow 0 \text{ as } n \to \infty.$$

By stationarity, this implies (F). Conversely, suppose that (F) holds, i.e. that $\mathbb{P}(F) = 1$. Since F is the intersection of the sets

$$F_n = \big\{\underline{\varphi} : \text{ there is } m \le n \text{ such that } |\mathrm{Im}\varphi_m^n| = 1\big\}$$

each of these sets has full measure 1 as well. Fix n now. Plainly, the sets

$$F_m^n = \big\{\underline{\varphi} : \ |\mathrm{Im}\varphi_m^n| = 1\big\}$$

increase to F_n as m decreases to $-\infty$. Hence there is $m < n$ such that $\mathbb{P}(F_m^n) > 0$. Choose now $x \neq y$ in $\boldsymbol{X}$. Since φ_1^{m-n+1} and φ_m^n are equal in law, for $\tau = n - m + 1$ one has

$$p^\tau\left(\{\varphi_1, \ldots, \varphi_\tau\} : \varphi_1^\tau(x) = \varphi_1^\tau(y)\right) = \mathbb{P}\big(\underline{\varphi} : \varphi_m^n(x) = \varphi_m^n(y)\big) \ge \mathbb{P}\big(F_m^n\big) > 0,$$

and (C) holds. $\square$

This shows that any derivation of coupling from the past which does not explicitly or implicitly use a hypothesis like (C) or a suitable substitute is necessarily incomplete or incorrect.

Remark It is tempting to transfer the same idea to 'coupling to the future'. Unfortunately, starting at zero and returning the first state of complete coalescence after zero, in general does not give a sample from μ.

The reader may want to check the following simple example from [5].

Example 4 Let $\boldsymbol{X} = \{1, 2\}$. Positive transition probabilities P and their invariant distributions μ have the form

$$P := \begin{pmatrix} 1-\lambda & \lambda \\ \kappa & 1-\kappa \end{pmatrix}, \ 0 < \lambda, \kappa < 1, \quad \mu = \left(\frac{\kappa}{\lambda+\kappa}, \frac{\lambda}{\lambda+\kappa}\right).$$

Start two independent chains η and ξ with transition probability P at time 0 from 1 and 2, respectively. The time of first coalescence in the future is

$$T := \min\{m \in \mathbb{N} : \eta_m = \xi_m\}.$$

Denote the common law of η_T and ξ_T by ϱ. We will shortly verify that $\varrho = \mu$ if and only if $\lambda = \kappa$. Compute first

$$\begin{aligned}
&\mathbb{P}\big(\eta_n = \xi_n = 1, \eta_m \neq \xi_m, m < n\big) \\
&\quad = \kappa(1-\lambda) \sum_{k=0}^{n} \tbinom{n}{k} \big((1-\lambda)(1-\kappa)\big)^k \big(\lambda\kappa\big)^{n-k} \\
&\quad = \kappa(1-\lambda)\big((1-\lambda)(1-\kappa) + \lambda\kappa\big)^n = \kappa(1-\lambda)\big(1 - (\lambda + \kappa - 2\lambda\kappa)\big)^n
\end{aligned}$$

and

$$\varrho(1) = \kappa(1-\lambda)\sum_{n=0}^{\infty}\left(1-(\lambda+\kappa-2\lambda\kappa)\right)^n = \frac{\kappa(1-\lambda)}{\kappa(1-\lambda)+\lambda(1-\kappa)}.$$

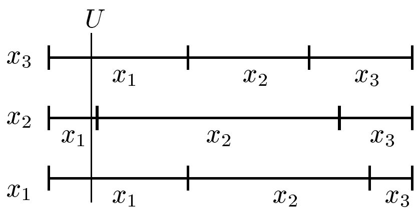
Hence

$$\varrho = \left(\frac{\kappa(1-\lambda)}{\kappa(1-\lambda)+\lambda(1-\kappa)}, \frac{\lambda(1-\kappa)}{\kappa(1-\lambda)+\lambda(1-\kappa)}\right).$$

This is the invariant distribution μ if and only if $\lambda = \kappa$.

The representation of Markov chains by stochastic flows is closely connected to the actual implementation of coupling from the past. Extending previous notation, a *transition rule* will be a map $f : \boldsymbol{X} \times \Theta \to \boldsymbol{X}$, with some set Θ to be specified. Let now V_i, $i \in \mathbb{Z}$, be independent identically distributed random variables taking values in Θ. Then $\varphi_i = f(\cdot, V_i)$, $i \in \mathbb{Z}$, is a stochastic flow. If, moreover, $\mathbb{P}(f(x, V_i) = y) = P(x, y)$ then the flow fulfills Condition (P). The remaining problem is to construct a transition rule such that the associated flow fulfills Condition (C) too.

Example 5 Recall from Example 1 how a Markov chain was realized there. Let again $f(x, u)$ be a deterministic transition rule taking values in $\boldsymbol{X}$, such that for a random variable U with uniform distribution on $\Theta = [0, 1]$ the variable $f(x, U)$ has law $P(x, \cdot)$. This way we - theoretically - may for an $m \le 0$ realize all values $\varphi_m^0(x)$, $x \in \boldsymbol{X}$, and check coalescence. If we go back k more steps in time we need all $\varphi_m^0 \circ \varphi_{m-k}^{m-1}(x)$. Since the maps $\varphi_0, \ldots, \varphi_m$ are kept, we must work with the same random numbers $u_0, \ldots, u_m$, i.e. realizations of the $U_0, \ldots, U_m$, as in the preceding run, and only independently generate additional random numbers $u_{m-1}, \ldots, u_{m-k}$. For this special coupling there is complete coalescence at time 0 in finite time. The strength of coupling depends on the special form of f which in turn depends on the concrete implementation.
In Example 1, for each $x \in X$, we partitioned $[0, 1]$ into intervals I_y^x of length $P(x, y)$ and in step n took that y with $U_n \in I_y^x$. The intervals $I_{y^*}^x$ with left end at 0 have an intersection I_{y^*} of length at least $\min_{x,y} P(x, y)$. This simultaneously is the probability that U falls into I_{y^*} and all states coalesce in y^* in one single step, irrespective of x. We may improve coupling by a clever arrangement of the intervals. If we put the intervals $I_{y^*}^x$ for which $\min\{|I_{y^*}^x| : x \in \boldsymbol{X}|\}$ is maximal, to the left end of $[0, 1]$ then we get the lower bound $\max_y \min_x P(x, y)$ for the coalescence probability. We can improve coupling even further, splitting the intervals into pieces of length $\min\{|I_y^x| : x \in \boldsymbol{X}\}$ and their rest, and arrange the equal pieces on the left of $[0, 1]$. This gives a bound $\sum_y \min_x P(x, y)$.



Note that although all these procedures *realize the same Markov kernel* P they correspond to *different transition rules*, to *different stochastic flows*, and to *different couplings*. Apart from all these modifications, we can summarize:

Proposition 1 *Suppose that $P > 0$. Then all stochastic flows $\varphi_i = f(\cdot, U_i)$ from the present Example 5 fulfill Condition* (C).

Note that the distribution of all these random maps definitely is not the synchronous one from Example 2. For this distribution, set $\Theta = [0,1]^{|\boldsymbol{X}|}$, use independent copies U_k^z, $z \in \boldsymbol{X}$, of U_k, and let $\varphi_k(x) = f(x, (U_k^z)_{z \in \boldsymbol{X}}) = g(x, U_k^x)$ for g on $\boldsymbol{X} \times [0,1]$ constructed like above. Condition (C) is obviously fulfilled and coupling from the past works also for this method.

Remark In Example 5 we found several lower bounds for the probability that states coalesce in one step. An upper bound is given by

$$\begin{aligned}\mathbb{P}\big(\varphi(x) = \varphi(y)\big) &= \sum_z \mathbb{P}\big(\varphi(x) = z, \varphi(y) = z\big) \\ &\le \sum_z \mathbb{P}\big(\varphi(x) = z\big) \wedge \mathbb{P}(\varphi(y) = z) = \sum_z P(x,z) \wedge P(y,z).\end{aligned}$$

This is closely related to DOBRUSHIN'S contraction technique, which in the finite case is based on *Dobrushin's contraction coefficient* $c(P) = 1 - \sum_z P(x,z) \wedge P(y,z)$, cf. [13], Chapter 4. The relation is

$$\mathbb{P}\big(\varphi(x) = \varphi(y)\big) \le 1 - c(P).$$

This upper bound is not sharp.

3 Monotonicity

Checking directly whether there is complete coalescence at time 0 starting at more and more remote past times and at all possible states is time consuming, and even impossible if the state space is large (as it is in the applications we have in mind). If coalescence of very few states enforces coalescence of all other states then the procedure becomes feasible. One of the concepts to make this precise is *monotonicity*. We are now going to introduce this concept on an elementary level.

Definition 1 *A **partial order** on a set $\boldsymbol{X}$ is a relation $x \preceq y$ between elements $x, y \in \boldsymbol{X}$ with the two properties*

(i) *$x \preceq x$ for each $x \in \boldsymbol{X}$ (**reflexivity**)*

(ii) *$x \preceq y$ and $y \preceq z$ implies $x \preceq z$ (**transitivity**).*

Recall that a *total order* requires the additional condition that any two elements $x, y \in \boldsymbol{X}$ are *comparable*, i.e $x \preceq y$ or $y \preceq x$.

Example 6 (a) The usual relation $x \le y$ on $\mathbb{R}$ is a *total order*. In the *componentwise order* on $\mathbb{R}^d$, $(x_1, \ldots, x_d) \preceq (y_1, \ldots, y_d)$ if and only if $x_i \le y_i$ for each i. It is a partial but no total order since elements like $(0,1)$ and (1,0) are not related.

(b) If $\boldsymbol{X} = \{\pm 1\}^S$, then in the component-wise order from (a), the constant configurations $b \equiv 1$ and $w \equiv -1$ are *maximal* and *minimal*, respectively, i.e. $x \preceq b$ and $w \preceq x$ for every $x \in \boldsymbol{X}$. This will be exploited in exact sampling for the Ising field in Section 4.

Next we want to lift partial orderings to the level of probability distributions. Call a subset I of $\boldsymbol{X}$ an *order ideal* if $x \in I$ and $y \preceq x$ imply $y \in I$.

Example 7 (a) The order ideals in $\mathbb{R}$ with the usual order are the rays $(-\infty, u]$ and $(-\infty, u)$, $u \in \mathbb{R}$.
(b) In the binary setting of Example 6(b), $x \preceq y$ if each black pixel of x is also black in y (if we agree that $x_s = +1$ means that the colour of pixel s is black). The order ideals are of the form $\{x \in \boldsymbol{X} : x \preceq y\}$.

Definition 2 *Let $(\boldsymbol{X}, \preceq)$ be a finite partially ordered set, and let ν and μ be probability distributions on $\boldsymbol{X}$. Then $\nu \preceq \mu$ in* ***stochastic order****, if and only if $\nu(I) \geq \mu(I)$ for each order ideal I.*

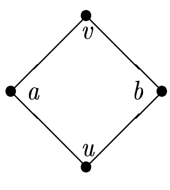
Example 8 Let ν and μ be distributions on $\mathbb{R}$ with cumulative distribution functions F_ν and F_μ, respectively. Then $\nu \preceq \mu$ if and only if $\mu((-\infty, u]) \leq \nu((-\infty, u])$ if and only if $F_\mu(u) \leq F_\nu(u)$ for every $u \in \mathbb{R}$. This means that 'the mass of ν is more on the left than the mass of μ'. For Dirac distributions $\varepsilon_u \preceq \varepsilon_v$ if and only if $u \leq v$.

The natural extension to Markov kernels reads

Definition 3 *We call a Markov kernel P on a partially ordered space $(\boldsymbol{X}, \preceq)$* ***stochastically monotone****, if and only if $P(x, \cdot) \preceq P(y, \cdot)$ whenever $x \preceq y$.*

In Example 5 we constructed transition rules f for homogeneous Markov chains, or rather Markov kernels P. A transition rule is called *monotone* if $f(x, u) \preceq f(y, u)$ for each u whenever $x \preceq y$. Plainly, a monotone transition rule induces a monotone Markov kernel. Conversely, a monotone kernel is not necessarily induced by a monotone transition rule, even in very simple situations. D.A. Ross (1993), see [3], p. 2., gives a simple counterexample:

Example 9 Consider the space $\boldsymbol{X} = \{u, v, a, b\}$ and let $u \preceq a, b$, and $a, b \preceq v$. Define a Markov kernel P by



$$P(u,u) = 1/2 = P(u,a), \quad P(a,u) = 1/2 = P(a,v)$$
$$P(b,a) = 1/2 = P(b,b), \quad P(v,a) = 1/2 = P(v,v)$$

The order ideals are $\emptyset$, $\{u\}$, $\{a, u\}$, $\{b, u\}$ and $\boldsymbol{X}$, and it is readily checked that P is monotone. Suppose now that there are random variables with $\xi_u \preceq \xi_a$, $\xi_b \preceq \xi_v$ almost surely and with laws $P(u, \cdot)$, $P(a, \cdot)$, $P(b, \cdot)$, and $P(v, \cdot)$, respectively. We shall argue that

$$\begin{aligned} \mathbb{P}(\xi_u = a) &= \mathbb{P}(\xi_u = a, \xi_a = v, \xi_b = a, \xi_v = v) = 1/2 \\ \mathbb{P}(\xi_b = b) &= \mathbb{P}(\xi_u = u, \xi_b = b, \xi_v = v) = 1/2. \end{aligned}$$

The two events are disjoint and hence $\mathbb{P}(\xi_v = v) = 1$ in contradiction to $\mathbb{P}(\xi_v = v) = 1/2$. We finally indicate how for example the first identity can be verified: Since $\xi_u \preceq \xi_a$ one has $\mathbb{P}(\xi_u = a) = \mathbb{P}(\xi_u = a, \xi_a \in \{a, v\})$. Since $\mathbb{P}(\xi_a = a) = 0$, we conclude $\mathbb{P}(\xi_u = a) = \mathbb{P}(\xi_u = a, \xi_a = v)$. Now repeat this argument two times.

Suppose now that the partially ordered space $(\boldsymbol{X}, \preceq)$ contains a *minimal* element u and a *maximal element* v, i.e. $u \preceq x \preceq v$ for every $x \in X$. Suppose further that the stochastic flow is induced by a monotone transition rule, i.e. $\varphi_i(x) = f(x, U_i)$ and $f(x, u) \preceq f(y, u)$ if $x \preceq y$. Then

$$\varphi_m^n(u) \preceq \varphi_m^n(x) \preceq \varphi_m^n(v) \text{ for every } x \in \boldsymbol{X},\ m \leq n,$$

and $\varphi_m^0(x) = w$, $m \leq 0$, for each $x \in \boldsymbol{X}$, as soon as $\varphi_m^0(u) = w = \varphi_m^0(v)$. The previous findings can be turned into practicable algorithms.

Proposition 2 *Suppose that P is monotone and $(\boldsymbol{X}, \preceq)$ has a minimum u and maximum v. Then coalescence for u and v enforces complete coalescence.*

4 Random Fields and the Ising Model

Random fields serve as flexible models in image analysis and spatial statistics. In particular, any full probabilistic model of textures with random fluctuations necessarily is a random field. Recursive (auto-associative) neural networks can be reinterpreted in this framework as well, cf. e.g. G. WINKLER (1995). To understand the phenomenology of these models, sampling from their *Gibbs distribution* provides an important tool. In the sequel we want to show how the concepts developed above serve to establish exact sampling from the Gibbs distribution of a well known random field – the Ising model.

Let a *pattern* or *configuration* be represented by an array $x = (x_s)_{s \in S}$ of 'intensities' $x_s \in G_s$ in 'pixels' or 'sites' $s \in S$ with finite sets G_s and S. S might be a finite square grid or - in case of neural networks - an undirected finite graph. A (finite) *random field* is a strictly positive probability measure Π on the space $\boldsymbol{X} = \prod_{s \in S} G_s$ of all configurations x. Taking logarithms shows that Π is of the *Gibbsian form*

$$\Pi(x) = Z^{-1} \exp(-K(x)),\ Z = \sum_z \exp(-K(z)), \tag{11}$$

with a function K on $\boldsymbol{X}$. It is called a *Gibbs fields* with *energy function* K and *partition function* Z. These names remind of their roots in statistical physics.

For convenience we restrict ourselves to the Gibbs sampler with random visiting scheme. Otherwise we had slightly to modify the setup of Section 2. Let pr_t be the projection $\boldsymbol{X} \to G_t$, $x \mapsto x_t$. For a Gibbs field Π let

$$\Pi\big(y_s \,\big|\, x_t, t \neq s\big) = \Pi\big(pr_s = y_s \,\big|\, pr_t = x_t, t \neq s\big) \tag{12}$$

denote the single-site conditional probabilities. The *Gibbs sampler with random visiting scheme* first picks a site $s \in S$ at random from a probability distribution D on S, and then picks an intensity at random from the conditional distribution (12) on G_s. Given a configuration $x = (x_t)$ this results in a new configuration $y = (y_t)$ which equals x everywhere except possibly at site s. The procedure is repeated with the *new* configuration y, and so on and so on. This defines a homogeneous Markov chain on $\boldsymbol{X}$ with Markov kernel

$$P(x,y) = \sum_{s \in S} D(s) \Pi_{\{s\}}(x,y), \ \ x, y \in \boldsymbol{X}, \tag{13}$$

where $\Pi_{\{s\}}(x,y) = \Pi\big(y_s \,\big|\, x_t, t \neq s\big)$ if x and y are equal off s and $\Pi_{\{s\}}(x,y) = 0$ otherwise. These transition probabilities $\Pi_{\{s\}}$ are called the *local characteristics*. D is called the *proposal* or *exploration distribution*.

We assume that D is strictly positive; frequently it is the uniform distribution on S. Then P is primitive since $P^{|S|}$ is strictly positive. In fact, in each step each site and each intensity in the site has positive probability to be chosen, and thus each y can be reached from each x in $|S|$ steps with positive probability. It is easily checked - verifying the detailed balance equations - that Π is the invariant distribution of P, and thus the invariant distribution of the homogeneous Markov chain generated by P.

Example 10 (The Ising model) Let us give an example for exact sampling by way of the Ising model. The *ferromagnetic Ising model with magnetic field* $h := (h_s)_{s \in S}$ is a binary random field with $G_s = \{-1, 1\}$ and energy function

$$K(x) = \beta \sum_{s \sim t} x_s x_t - \sum_s h_s x_s,$$

where $\beta > 0$, $h_s \in \mathbb{R}$ and $s \sim t$ indicates that s and t are neighbours. For the random visiting scheme in (13) the Markov chain is homogeneous and fits perfectly into the setting of Section 2. The formula from [13], Proposition 3.2.1 (see also [13], Example 3.1.1) for the local characteristics boils down to

$$p^+(x) = \Pi\big(X_s = 1 \,\big|X_t = x_t, \ t \neq s\big) = \Big(1 + \exp\big(-2\beta \sum_{t \sim s} x_t - h_s\big)\Big)^{-1}.$$

This probability increases with the set $\{t \in S : x_t = 1\}$. Hence $p^+(y) \geq p^+(x)$ if $x \preceq y$ in the component-wise partial order introduced in Example 6. The updates x' and y' preserve all the black sites off s, and possibly create an additional black one at s. We conclude that P from (13) is monotone and fulfills the hypotheses of Proposition 2. Hence for complete coalescence one only has to check whether the completely black and the completely white patterns coalesce. For transition rules like in Example 5 the Condition (C) on page 9 is also fulfilled and coupling from the past works.

5 Conclusion

The authors are not aware of other mathematical fields, where so many insufficient arguments, ranging from incomplete or misleading, to completely wrong, have been published (mainly in the Internet). In particular, Condition (C) or a substitute for it, are missing in a lot of presently available texts. A rigorous treatment is S.G. Foss and R.L. Tweedie (1998). These authors do not use iterated random maps. These are exploited systematically in P. Diaconis and D. Freedman (1999). J.A. Fill (1998) introduces 'interruptible' perfect sampling based on acceptance/rejection sampling. Meanwhile there is a body of papers on exact sampling. On the other hand, the field still is in the state of flux and hence it does not make sense to give further references; a rich and up to date source is the home-page of D.B. Wilson, `http://www.dbwilson.com/exact/`. The connection between transition probabilities and random maps was clarified in H.v. Weizsäcker (1974).

Acknowledgement: We thank H.v. Weizsäcker, Kaiserslautern, for helpful discussions during the initial phase of the work.

References

[1] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Rev.*, 41 (1):45–76, 1999.

[2] J.A. Fill. An interuptible algorithm for perfect sampling via Markov chains. *The Ann. of Appl. Probab.*, 8(1):131–162, 1998.

[3] J.A. Fill and M. Machida. Stochastic monotonicity and realizable monotonicity. *Ann. Probab.*, 29:938–978, 2001.

[4] S.G. Foss and R.L. Tweedie. Perfect simulation and backward coupling. *Stoch. Models*, 14(1-2):187–204, 1998.

[5] F. Friedrich. *Sampling and statistical inference for Gibbs fields*. PhD thesis, University of Heidelberg, Munich, Germany, 2003. draft.

[6] A. Gelman. Inference and monitoring convergence. In W.R. Gilks et al. (1996b), chapter 8, pages 131–143.

[7] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. Introducing Markov chain Monte Carlo. In W.R. Gilks et al. (1996b), chapter 1, pages 1–19.

[8] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Interdisciplinary Statistics. Chapman & Hall, London, Weinheim, New York, Tokyo, Melbourne, Madras, 1996b.

[9] J.G. Propp and D.B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

[10] A.E. Raftery and S.M. Lewis. Implementing MCMC. In W.R. Gilks et al. (1996b), chapter 7, pages 115–130.

[11] D.A. Ross. A coherence theorem for ordered families of probability measures on a partially ordered space. Unpublished manuscript, 1993.

[12] H.v. Weizsäcker. Zur Gleichwertigkeit zweier Arten der Randomisierung. *Manuscripta Mathematika*, 11:91–94, 1974.

[13] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, volume 27 of *Applications of Mathematics*. Springer Verlag, Berlin, Heidelberg, New York, 1995.

[14] G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, volume 27 of *Applications of Mathematics*. Springer Verlag, Berlin, Heidelberg, New York, second edition, 2003.