Problem solving is often a matter of cooking up an appropriate Markov chain

Olle Häggström*

January 15, 2007

Abstract

By means of a series of examples, taken from classic contributions to probability theory as well as from the author's own practice, an attempt is made to convince the reader that *problem solving is often a matter of cooking up an appropriate Markov chain.* Topics touched upon along the way include coupling, correlation inequalities, and percolation.

Key words: correlation inequality, coupling, Markov chain, MCMC, percolation, stochastic domination.

Running head: Problem solving using Markov chains.

1 Introduction

A century has passed since the introduction by A.A. Markov of what we now know as Markov chains; see Markov (1906) and Basharian et al. (2004). During that period, Markov chains have turned out to be not only a rich source of beautiful mathematics, but also immensely useful in a variety of applied areas such as statistical mechanics, queueing theory, information theory, statistics, speech recognition, and bioinformatics, just to name a few.

The most common way to use Markov chains in these and other areas is as ingredients in the modelling of one kind or another of time dynamics. A completely different use of Markov chains is the so-called Markov chain Monte Carlo (MCMC) method, pioneered by Metropolis et al. (1953), Hastings

*Chalmers University of Technology, Sweden

(1970), Geman and Geman (1984) and others. Here, Markov chains are applied to situations that in themselves need not involve any time dynamics at all. The problem is to generate computer samples with some prescribed but typically very complicated distribution π on some large state space S, and the idea of MCMC is that in situations where it appears practically impossible to sample directly from π, it may be easy to sample from the transition kernel of some irreducible and aperiodic Markov chain $X = \{X(0), X(1), \ldots\}$ on S whose unique stationary distribution is precisely π. If the chain has the property of rapid convergence to stationarity (as hopefully it has), then an easy way to generate an S-valued random object whose distribution is close to π, is to start the chain with $X(0)$ chosen arbitrarily, to run it for a while (say, time n), and output $X(n)$. See, e.g., Gilks et al. (1996) or Häggström (2002) for introductions to the theory and practice of MCMC.

The purpose of the present paper is to elaborate on the somewhat less well-known idea that the central ingredient in MCMC – namely the introduction of a Markov chain designed to have a prescribed stationary distribution π – is useful in a variety of contexts that do not involve computer simulations of any sort. Rather, in the kind of applications I have in mind, it is not necessary to implement and run the Markov chains: it will suffice to think about them on a more abstract level. Every mathematician needs to have a toolbox of devices and tricks to use in various situations, and I hope to convince readers that the readiness to try out such Markov chain ideas is a useful enough device that they will want to include it in their own toolboxes.

At this point, a line from Lindvall's (1992) influential introduction to coupling methods seems apt: "To know a method is to have learned how it works. What we have ahead of us is a collection of applications of a few basic ideas" (p. 6). In the remaining sections, I will focus on three basic examples. In Section 2 I will discuss the use of Markov chains for proving the very useful correlation inequality of Harris (1960). Then, I will go on to discuss examples from my own practice: in Section 3 a domination inequality needed in a problem arising in survey sampling, and in Section 4 a conditional correlation inequality for percolation models.

One aspect of my Lindvallian approach in this paper is that I have no pretentions of providing an exhaustive survey of the topic. For a particular subtopic which is left out of the discussion but which I recommend to the ambitious reader, let me mention the exploitation of ideas from the coupling-from-the-past approach of Propp and Wilson (1996) in so-called perfect MCMC, to the rigorous analysis of ergodic properties of Gibbisan random fields; this idea was first conceived by van den Berg and Steif (1998) and then further exploited in Häggström and Steif (2000), Häggström et al.

(2000) and Häggström et al. (2002).

2 Harris' inequality

Harris (1960) is a classic paper. Way ahead of its time, it contains a number of ideas that have influenced percolation theory for decades, and one – a correlation inequality now known as Harris' inequality – whose influence has extended far beyond that field.

Percolation theory (see Grimmett (1999) for an introduction) deals with connectivity properties of random media, and the basic mathematical model is as follows. Let $G = (V, E)$ be a (finite or countably infinite) connected graph with vertex set V and edge set E, where each edge $e \in E$ links two of the vertices. When G is infinite it is customary to also impose the condition of local finiteness, meaning that each $x \in V$ is incident to only finitely many edges. A main example is to let G be the infinite square lattice, which is denoted $\mathbf{Z}^2$ and which arises by letting V consist of all integer points in the Euclidean plane and having edges between vertices at Euclidean distance 1 from each other. The so-called retention parameter $p \in [0, 1]$ is fixed, and each edge in G is removed independently with probability $1 - p$. This produces a random subgraph of G, and the percolation-theoretic challenge is to say something about the connected components of this subgraph. For instance, if G is infinite, one may ask whether an infinite connected component occurs. The probability of getting an infinite connected component is easily shown to be 0 or 1 depending on whether p is above or below a critical value $p_c = p_c(G) \in [0, 1]$. The main result of Harris (1960) was that for the square lattice, $p_c \geq \frac{1}{2}$. This inequality was conjectured to be in fact an equality, but it took another 20 years before that was rigorously estabished by Kesten (1980).

What I've described here is so-called bond percolation, as opposed to site percolation which has much the same flavor but where it is vertices rather than edges that are removed at random.

Let me give a vague motivation for why correlation inequalities are important in percolation theory. Establishing connectivities over long distances often proceeds through a kind of concatenation procedure. For two vertices $x, y \in V$, an obvious sufficient condition for the existence of a path between x and y – an event that we denote by $\{x \leftrightarrow y\}$ – is that both of them have paths to some third vertex $z \in V$. Thus,

$$\mathbf{P}(x \leftrightarrow y) \;\geq\; \mathbf{P}(x \leftrightarrow z, z \leftrightarrow y)\,.$$

Here it is typically useful to be able to conclude that

$$\mathbf{P}(x \leftrightarrow y) \geq \mathbf{P}(x \leftrightarrow z)\mathbf{P}(z \leftrightarrow y)\,,$$

an argument that however requires that the events $\{x \leftrightarrow z\}$ and $\{z \leftrightarrow y\}$ are positively correlated in the sense that

$$\mathbf{P}(x \leftrightarrow z, z \leftrightarrow y) \geq \mathbf{P}(x \leftrightarrow z)\mathbf{P}(z \leftrightarrow y)\,. \tag{1}$$

Are they? The answer is yes, by an application of Harris' inequality (Theorem 1 below).

To set the stage for this result, we need some definitions. We will be concerned with a collection $\{X_i\}_{i\in I}$ of real-valued random variables, where the index set I is always taken to be finite or countably infinite. For $x, x' \in \mathbf{R}^I$, we write $x \preceq x'$ if $x_i \leq x'_i$ for every $i \in I$, and we call a function $f : \mathbf{R}^I \to \mathbf{R}$ increasing if $f(x) \leq f(x')$ whenever $x \preceq x'$.

Theorem 1 (Harris' inequality) *Let $X = \{X_i\}_{i\in I}$ be a collection of independent real-valued random variables, and let $f, g : \mathbf{R}^I \to \mathbf{R}$ be two bounded and increasing functions. Then*

$$\mathbf{E}[f(X)g(X)] \geq \mathbf{E}[f(X)]\mathbf{E}[g(X)]\,. \tag{2}$$

The significant condition here on f and g is that they are increasing, while the boundedness is just a convenient way of making sure that the expectations in (2) exist. The property that (2) holds for all bounded and increasing f and g is sometimes known as the positive associations property of $\{X_i\}_{i\in I}$, so with this terminology Harris' inequality says that any collection of independent real-valued random variables is also positively associated.

To see how (1) follows from Harris' inequality, we first equip each edge $e \in E$ with a random variable X_e taking value 1 or 0 depending on whether the edge e is present or not after the random thinning of the graph G. That makes $\{X_e\}_{e\in E}$ a collection of i.i.d. Bernoulli (p) random variables. The indicator function $\mathbf{1}_{\{x\leftrightarrow z\}}$ is increasing in these variables, because increasing the X_e's means inserting edges, and inserting edges cannot take us out of the event $\{x \leftrightarrow z\}$. The same goes for the indicator $\mathbf{1}_{\{z\leftrightarrow y\}}$, so Harris' inequality tells us in particular that

$$\mathbf{E}[\mathbf{1}_{\{x\leftrightarrow z\}}\mathbf{1}_{\{z\leftrightarrow y\}}] \geq \mathbf{E}[\mathbf{1}_{\{x\leftrightarrow z\}}]\mathbf{E}[\mathbf{1}_{\{z\leftrightarrow y\}}]$$

which is just another way of expressing (1).

There are various ways to prove Theorem 1 – see Harris (1960) or Grimmett (1999) – besides the Markov chain approach employed here which I

personally find the most illuminating. This approach goes back to Holley (1974). The core of the matter lies in proving the following special case consisting of a finite collection of i.i.d. $\{0,1\}$-valued random variables; once that is done, the general case follows, as we shall see, in fairly straightforward manner.

Proposition 1 *Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli (p) random variables, and let $f, g : \{0,1\}^n \to \mathbf{R}$ be increasing functions. Then*

$$\mathbf{E}[f(X_1, \ldots, X_n)g(X_1, \ldots, X_n)] \geq \mathbf{E}[f(X_1, \ldots, X_n)]\mathbf{E}[g(X_1, \ldots, X_n)]\,. \tag{3}$$

A key ingredient in the preferred proof of this result, besides Markov chains, is the notion of a coupling. A coupling of two probability distributions μ and μ' is a joint construction on the same probability space of two random objects with respective distributions μ and μ' done with the explicit purpose of drawing conclusions about (and sometimes comparing) these distributions. The coupling idea is best explained via examples, as will be done below, but see also Lindvall (1992) and Thorisson (1995, 2000) for introductions to coupling methods.

Here, let μ denote the probability distribution on $\{0,1\}^n$ of $(X_1, \ldots, X_n)$, as above i.i.d. Bernoulli (p). Furthermore, let μ_g be the so-called g-biased perturbation of μ, defined by setting

$$\mu_g(\omega) \;=\; Z^{-1}\mu(\omega)g(\omega)$$

for each $\omega \in \{0,1\}^n$, where

$$Z = \sum_{\omega \in \{0,1\}^n} \mu(\omega)g(\omega)$$

is a normalizing constant. Of course, this makes μ_g a probability measure only if g is nonnegative (and not identically zero). But since adding a constant to the function g doesn't change whether or not (3) holds, there is no loss of generality in assuming that $g(\omega) > 0$ for all $\omega \in \{0,1\}^n$. So let us assume that.

An intermediate step in proving Proposition 1 is the following lemma.

Lemma 1 *It is possible to couple two $\{0,1\}^n$-valued random variables X and Y with respective distributions μ and μ_g, such that*

$$\mathbf{P}(X \preceq Y) \;=\; 1\,. \tag{4}$$

Note in particular that since f is increasing, (4) implies that $\mathbf{E}[f(X)] \leq \mathbf{E}[f(Y)]$. Once we have the lemma, Proposition 1 follows by a simple calculation. Note first that in terms of μ, the desired inequality (3) may be written as

$$\mu(fg) \geq \mu(f)\mu(g)\,, \tag{5}$$

and that $Z = \mu(g)$. We get, with X and Y as in Lemma 1,

$$\begin{aligned}
\mu(f) = \mathbf{E}[f(X)] &\leq \mathbf{E}[f(Y)] \\
&= \mu_g(f) \\
&= \sum_{\omega\in\{0,1\}^n} \mu_g(\omega) f(\omega) \\
&= \frac{\sum_{\omega\in\{0,1\}^n} \mu(\omega) g(\omega) f(\omega)}{Z} \\
&= \frac{\mu(fg)}{\mu(g)}
\end{aligned}$$

and multiplying by $\mu(g)$ yields (5). Thus, in order to prove Proposition 1, it only remains to prove Lemma 1.

Proof of Lemma 1: Here is where the long-awaited Markov chains enter our game. We will begin by defining two $\{0,1\}^n$-valued Markov chains $(X(0), X(1), X(2), \ldots)$ and $(Y(0), Y(1), Y(2), \ldots)$ designed to have μ and μ_g as their respective unique stationary distributions.

The transition mechanism for $(X(0), X(1), X(2), \ldots)$ is as follows. Given $X(k)$, set $i = k(\bmod n) + 1$ and set, independently of everything else,

$$X_i(k+1) = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases}$$

while setting $X_j(k+1) = X_j(k)$ for all $j \neq i$. (Note that this makes the chain time-inhomogeneous with a transition kernel that repeats itself every n time units.)

It is obvious that if $X(k)$ has distribution μ, then so has $X(k+1)$. So μ is a stationary distribution for the chain. And it is equally obvious that the chain is irreducible and aperiodic, so that $X(k)$ converges in distribution to μ as $k \to \infty$ no matter how it is started. (Note: While it is true that in the time-inhomogeneous case irreducibility and aperiodicity are not in general sufficient for a finite-state Markov chain to exhibit such convergence, here this is not a problem because sampling X at every n'th time gives a time-homogeneous chain with the corresponding properties.) Readers familiar

with MCMC will notice that this Markov chain is precisely the so-called systematic scan Gibbs sampler for the distribution μ.

Let us construct $(Y(0), Y(1), Y(2), \ldots)$ in the same vein. To this end, for $i \in \{1, \ldots, n\}$ and $\xi \in \{0,1\}^{\{1,\ldots,n\}\setminus\{i\}}$ define $\gamma_{i,\xi}$ to be the conditional probability that the i:th coordinate of a $\{0,1\}^n$-valued random object with distribution μ_g takes value 1 given that the other coordinates are given by ξ. Let the transition mechanism for $(Y(0), Y(1), Y(2), \ldots)$ be as follows. Given $Y(k)$, we set $i = k(\bmod n) + 1$ and set

$$Y_i(k+1) = \begin{cases} 1 & \text{w.p. } \gamma_{i,\xi} \\ 0 & \text{w.p. } 1 - \gamma_{i,\xi} \end{cases}$$

where ξ is given by the values of $Y(k)$ on $\{1, \ldots, n\} \setminus \{i\}$; and finally we set $Y_j(k+1) = Y_j(k)$ for all $j \neq i$.

This makes $(Y(0), Y(1), Y(2), \ldots)$ another instance of the Gibbs sampler, irreducible and aperiodic, with $Y(k)$ converging in distribution to the chain's unique stationary distribution μ_g.

Next, we specify how to run the two chains simultaneously on the same probability space. We start the chains by picking $X(0)$ and $Y(0)$ indepedently according to their respective stationary distributions μ and μ_g. Let $(U_0, U_1, U_2, \ldots)$ be a sequence of i.i.d. random variables uniformly distributed on $[0,1]$. To go from time k to time $k+1$, set

$$X_i(k+1) = \begin{cases} 1 & \text{if } U_k < p \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and

$$Y_i(k+1) = \begin{cases} 1 & \text{if } U_k < \gamma_{i,\xi} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

I now claim, crucially, that

$$\gamma_{i,\xi} \geq p \tag{8}$$

regardless of i and ξ. This implies that as soon as a given coordinate i is chosen, we have $X_i(k) \leq Y_i(k)$ from that time k and forever after. As soon as all coordinates have been visited, which happens at time n, we thus have $X_i(n) \leq Y_i(n)$ for all i. Thus, picking the pair (X, Y) according to the joint distribution of $X(n)$ and $Y(n)$ produces the desired coupling that establishes the lemma.

It only remains to prove the claim (8). Write $\xi \vee 0$ (resp. $\xi \vee 1$) for the element of $\{0,1\}^n$ that equals 0 (resp. 1) at the i:th coordinate, and agrees with ξ elsewhere. Showing (8) is the same as showing that

$$\frac{\gamma_{i,\xi}}{1-\gamma_{i,\xi}} \geq \frac{p}{1-p}. \tag{9}$$

We get

$$\begin{aligned} \frac{\gamma_{i,\xi}}{1-\gamma_{i,\xi}} = \frac{\mu_g(\xi \vee 1)}{\mu_g(\xi \vee 0)} &= \frac{\mu(\xi \vee 1)g(\xi \vee 1)}{\mu(\xi \vee 0)g(\xi \vee 0)} \\ &\geq \frac{\mu(\xi \vee 1)}{\mu(\xi \vee 0)} \\ &= \frac{p}{1-p}, \end{aligned}$$

where the inequality is due to g being increasing. $\diamond$

Lemma 1 and, consequently, Proposition 1 are thus established.

It is a slightly unusual feature of this particular Markov chain proof that we got away with looking at the chains at a fixed finite time n; in the following two sections we will have to consider asymptotics as time tends to infinity.

Equipped with Proposition 1, we are now in a position to obtain Theorem 1 at low cost.

Proof of Theorem 1: We proceed by extending Proposition 1 via a couple of intermediate levels of generality. As a first step, consider the case where we allow an infinite collection $X = (X_1, X_2, \ldots)$ of i.i.d. variables, but still insist that they are binary. Define

$$f_n(X) = \mathbf{E}[f(X) \,|\, X_1, \ldots, X_n]$$

and $g_n(X)$ analogously. Both f_n and g_n are increasing, so Propsition 1 yields

$$\mathbf{E}[f_n(X)g_n(X)] \geq \mathbf{E}[f_n(X)]\mathbf{E}[g_n(X)]. \tag{10}$$

Furthermore, a standard application (to be found, e.g., in Kallenberg (1997), Thm. 6.23) of the martingale convergence theorem tells us that, a.s., $f_n(X) \to f(X)$ and $g_n(X) \to g(X)$ as $n \to \infty$. Thus, we may take limits in (10) to conclude that $\mathbf{E}[f(X)g(X)] \geq \mathbf{E}[f(X)]\mathbf{E}[g(X)]$.

As a next step, note that we can go directly from the case of infinitely many binary variables to that of finitely or infinitely many variables whose

distribution is uniform on $[0,1]$, simply by representing the latter by their binary expansions. (If $f : [0,1] \to \mathbf{R}$ is increasing in the usual sense, then $f(x)$ is also an increasing function of the binary expansion of x, while the converse is not true.)

Finally, to go from uniform $[0,1]$ variables to arbitrary real-valued random variables, it suffices to recall the inverse probability transform which tells us that any real-valued random variable can be obtained as a monotone transformation of a uniform $[0,1]$ random variable, while noting that the composition of two increasing functions is increasing. Theorem 1 is therefore established. $\diamond$

Extensions of Harris' inequality to certain classes of dependent random variables have been made. One contribution worth mentioning in this context is Esary et al. (1967). Arguably the most famous extension is the so-called FKG inequality of Fortuin, Kasteleyn and Ginibre (1971); see also Holley (1974) and Georgii et al. (2001). Here I feel compelled to point out that it is fairly common in the literature that alleged applications of the FKG inequality concern i.i.d. systems, so that in fact a lot of credit that rightfully should go to Harris ends up instead with the FKG trio.

Inspecting the Markov chain argument in the proof of Proposition 1 to see what assumptions on μ are really needed, and considering the asymptotic joint distribution of $(X(k), Y(k))$ as $k \to \infty$, leads to the variation of the FKG inequality that appears, e.g., in Thm. 4.11 of Georgii et al. (2001). Besides a technical assumption such as requiring that μ assigns positive probability to all $\omega \in \{0,1\}^n$ (this may be weakened), the crucial assumption is that for any i, the conditional μ-probability of seeing a 1 at coordinate i, given that the other coordiantes take values according to $\xi \in \{0,1\}^{\{1,\ldots,n\}\setminus\{i\}}$, is increasing as a function of ξ. This turns out to hold for many important examples, such as in the ferromagnetic Ising model and in the most relevant parts of the parameter space of the so-called random-cluster model; see Georgii et al. (2001) again.

3 A domination result for sampling

In 1995, I was approached by two of my local colleagues at Chalmers, Johan Jonasson and Olle Nerman, who were stuck on a seemingly obvious inequality which they needed in the context of survey sampling with unequal probabilities. Thanks to my knowledge of the Markov chain approach to Harris' inequality, I was quickly able to help them.

Fix n and $p_1, \ldots, p_n \in [0,1]$, and let $X_1, \ldots, X_n$ be independent (though

not necessarily identically distributed) Bernoulli variables with parameters $p_1, \ldots, p_n$. Write $S = \sum_{j=1}^{n} X_j$ for their sum. The question Jonasson and Nerman asked is whether, for any $i \in \{1, \ldots, n\}$ and $s \in \{0, \ldots, n-1\}$, it is the case that $\mathbf{P}(X_i = 1 \,|\, S = s) \leq \mathbf{P}(X_i = 1 \,|\, S = s+1)$. Intuitively this seems highly plausible: the larger S is, the more likely should any of the Bernoulli variables be to take value 1. And, yes:

Proposition 2 *With $X_1, \ldots, X_n$ and S as above, we have, for any $i \in \{1, \ldots, n\}$ and $s \in \{0, \ldots, n-1\}$ that*

$$\mathbf{P}(X_i = 1 \,|\, S = s) \;\leq\; \mathbf{P}(X_i = 1 \,|\, S = s+1)\,. \tag{11}$$

The similar inequality $\mathbf{P}(X_i = 1 \,|\, S \leq s) \leq \mathbf{P}(X_i = 1 \,|\, S > s)$ follows immediately from Harris' inequality (with $f(X) = X_i$ and $g(X) = \mathbf{1}_{\{S>s\}}$), but (11) requires a different proof. The following is how I argued using Markov chains.

Proof of Proposition 2: For $s = 0, \ldots, n$, let μ_s denote the probability measure on $\{0,1\}^n$ obtained by conditioning $(X_1, \ldots, X_n)$ on the event $\{S = s\}$. For each μ_s, we would like to devise a Markov chain with μ_s as its stationary distribution. Directly copying the Gibbs sampler approach in Section 2 won't do, because the μ_s-conditional probability of having a 1 at coordinate i given the values at all other coordinates is always degenerate, thus resulting in a boring Markov chain that ends up mapping any state onto itself with probability 1.

Instead, let us try a variant of the Gibbs sampler where we update *two* coordinates at a time. For fixed distinct $i, j \in \{1, ..., n\}$, the conditional distribution of (X_i, X_j) given the event $X_i + X_j = 1$, is given by

$$(X_i, X_j) = \begin{cases} (1,0) & \text{w.p. } \frac{p_i(1-p_j)}{p_i(1-p_j)+p_j(1-p_i)} \\ (0,1) & \text{w.p. } \frac{p_j(1-p_i)}{p_i(1-p_j)+p_j(1-p_i)}\,, \end{cases}$$

and this conditional distribution is unaffected by further conditioning on S. Therefore, $\mu_0, \mu_1, \ldots, \mu_n$ are all stationary distributions for the $\{0,1\}^n$-valued Markov chain $(X(0), X(1), \ldots)$ with the transition mechanism where at each time k we do the following:

1. Pick two indices $i, j \in \{1, \ldots, n\}$ at random according to uniform distribution without replacement.

2. If $X_i(k) = X_j(k)$, then set $X_i(k+1) = X_j(k+1) = X_i(k)$. Otherwise, set

$$(X_i(k+1), X_j(k+1)) = \begin{cases} (1,0) & \text{w.p. } \frac{p_i(1-p_j)}{p_i(1-p_j)+p_j(1-p_i)} \\ (0,1) & \text{w.p. } \frac{p_j(1-p_i)}{p_i(1-p_j)+p_j(1-p_i)}, \end{cases} \tag{12}$$

3. Set $X_h(k+1) = X_h(k)$ for all $h \notin \{i, j\}$.

Now let us run $(X(0), X(1), \ldots)$ together with a second $\{0,1\}^n$-valued Markov chain $(Y(0), Y(1), \ldots)$ with exactly the same transition kernel. We "synchronize" the transitions by

(a) always picking the same coordinates i and j in the Y chain as in the X chain, and

(b) whenever $X_i(k)+X_j(k) = Y_i(k)+Y_j(k) = 1$, we take $(X_i(k+1), X_j(k+1))$ and $(Y_i(k+1), Y_j(k+1))$ to be equal (and equal to $(1,0)$ or $(0,1)$ with the probabilities prescribed in (12)).

Now start the chains with $X(0)$ chosen according to μ_s, and independently $Y(0)$ according to μ_{s+1}.

Define, for each k, $Z(k)$ as the number of coordinates in which $X(k)$ and $Y(k)$ differ. Note that $Z(k) \geq 1$, with equality if and only if $X(k) \preceq Y(k)$. Furthermore, with the above synchronization of the two chains, we see that $(Z(0), Z(1), Z(2), \ldots)$ is a decreasing process, because $Z(k+1)$ will equal $Z(k)$ in all cases except when i and j happen to be chosen in such a way that $(X_i(k), X_j(k)) = (1,0)$ and $(Y_i(k), Y_j(k)) = (0,1)$ or vice versa, in which case we get $Z(k+1) = Z(k) - 2$. Whenever $Z(k) > 1$ there is a positive probability (bounded below by $2/n(n-1)$) that such i and j are chosen, and repeated application of the Borel–Cantelli lemma implies that a.s. the Z process keeps decreasing until eventually it reaches the absorbing level 1.

Hence, from some (random) time and onwards we will have $X(k) \preceq Y(k)$ and in particular that $X_i(k) \leq Y_i(k)$. By picking two $\{0,1\}^n$-valued random objects X and Y according to the asymptotic distribution as $k \to \infty$ of $(X(k), Y(k))$ (after passing to a subsequence if necessary – which incidentally it isn't), and recalling that $X(k)$ and $Y(k)$ have distributions μ_s and μ_{s+1} for each k (and thus also in the limit), we obtain (11). $\diamond$

This argument has to my knowledge previously appeared only in the preprint by Jonasson and Nerman (1996). Publication of their paper was delayed, and when eventually its descendant, Aires et al. (2002), appeared the work had evolved to the point where Proposition 2 was no longer needed.

Some years after Jonasson's and Nerman's original query, Yuval Peres and independently Tue Tjur explained to me, in response to my Markov chain argument, that Proposition 2 is in fact intimately related to an inequality of no less a soul than Isaac Newton.

In its simplest form, Newton's inequality states that if a polynomial

$$P(x) \;=\; a_0 + a_1 x + \ldots, + a_n x^n$$

with real coefficients has only real roots, then the coefficients $a_0, \ldots, a_n$ satisfy the log-concavity relation

$$a_{j-1} a_{j+1} \leq a_j^2 \tag{13}$$

for every j; see Newton (1707) or Niculescu (2000). The connection between this result and sums of Bernoulli variables proceeds via the observation that if $X_1, \ldots, X_n$ are independent Bernoulli variables with parameters $p_1, \ldots, p_n$, and S is their sum, then $\mathbf{P}(S = i)$ equals the coefficient a_i in the polynomial

$$\prod_{j=1}^{n} (1 - p_j + p_j x)\,.$$

Now, for $i \in \{1, \ldots, n\}$, define $S_i' = S - X_i$. Since S_i' is a sum of $n-1$ independent Bernoulli variables, we get from (13) that

$$\mathbf{P}(S_i' = s-1)\mathbf{P}(S_i' = s+1) \;\leq\; \mathbf{P}(S_i' = s)^2 \tag{14}$$

for any s. The inequality (11) in Proposition 2 is the same as saying that

$$\frac{\mathbf{P}(X_i = 1 \,|\, S = s)}{\mathbf{P}(X_i = 0 \,|\, S = s)} \;\leq\; \frac{\mathbf{P}(X_i = 1 \,|\, S = s+1)}{\mathbf{P}(X_i = 0 \,|\, S = s+1)}\,,$$

which, by rewriting the two ratios, we see is equivalent to

$$\frac{\mathbf{P}(X_i = 1, S_i' = s-1)}{\mathbf{P}(X_i = 0, S_i' = s)} \leq \frac{\mathbf{P}(X_i = 1, S_i' = s)}{\mathbf{P}(X_i = 0, S_i' = s+1)}\,. \tag{15}$$

The probability in the numerator of the left-hand side factors into $p_i \mathbf{P}(S_i' = s-1)$, and similarly for the three other probabilities, so (15) is the same as

$$\frac{p_i \mathbf{P}(S_i' = s-1)}{(1-p_i)\mathbf{P}(S_i' = s)} \leq \frac{p_i \mathbf{P}(S_i' = s)}{(1-p_i)\mathbf{P}(S_i' = s+1)}\,,$$

which of course follows via (14) from Newton's inequality.

Note that this reasoning can be turned around to derive Newton's inequality – at least in the case where all roots are negative – from Proposition 2. At one point, I therefore toyed with the idea of publishing a note with a title like "A probabilistic proof of an inequality of Newton", but decided against it, as (13) admits a relatively straightforward proof by induction in n (I leave this to the reader).

4 Conditioning and correlation in percolation

Harris' inequality and related results have proved extremely useful in percolation theory and related topics – see, e.g., Grimmett (1999) and Georgii et al. (2001) – a fact that motivates a considerable interest in trying to come up with new correlation inequalities. An important example is the so-called BK inequality of van den Berg and Kesten (1985) for "disjoint occurence" of increasing events, and the extension of this by Reimer (2000) to arbitrary events. Here let us look in another direction, namely that of whether (variants of) the Harris and FKG inequalities are preserved under various kinds of conditioning.

Let us focus on the standard bond percolation model on a finite or infinite but locally finite graph $G = (V, E)$ and retention parameter p, where each edge is independently deleted with probability $1 - p$, as in Section 2. Write $X \in \{0,1\}^E$ for the resulting random subgraph as represented by the indicator variables $X_e = \mathbf{1}_{\{e \text{ is retained}\}}$ for each $e \in E$. Harris' inequality tells us that for any two bounded and increasing functions $f, g : \{0,1\}^E \to \mathbf{R}$, we have

$$\mathbf{E}[f(X)g(X)] \geq \mathbf{E}[f(X)]\mathbf{E}[g(X)]\,.$$

Suppose now that we condition on an event $A \subset \{0,1\}^E$. Does the inequality

$$\mathbf{E}[f(X)g(X) \mid A] \geq \mathbf{E}[f(X) \mid A]\mathbf{E}[g(X) \mid A] \tag{16}$$

then hold? It is easy to devise examples that show that the answer is no, and that we do not even recover (16) by requiring that (the indicator function of) A is increasing or that it is decreasing.

But all is not lost. Restricting to certain "connectivity events" in $\{0,1\}^n$ allows us to derive certain correlation inequalities. For a vertex $x \in V$, write $\mathcal{F}_x$ for the σ-field consisting of events whose occurence or non-occurence can be determined from knowledge just of which vertices and which edges are part of the connected component of X containing x. Examples of events in $\mathcal{F}$ are

$$A = \{\text{the connected component containing } x \text{ has at least 10 edges}\}$$

and, for fixed $y \in V$, the event $B = \{x \leftrightarrow y\}$ that y is in the same connected component of X as x is. We write $\{x \not\leftrightarrow y\}$ for the complement of the latter event, and note that of course also this complement is in $\mathcal{F}_x$.

The following conditional correlation inequality was recently established by van den Berg et al. (2006a).

Theorem 2 *Consider bond percolation on a locally finite graph $G = (V, E)$ with retention parameter $p \in [0, 1]$. Then, for any two vertices $x, y \in V$ and any two bounded and increasing events $A \in \mathcal{F}_x$ and $B \in \mathcal{F}_y$, we have*

$$\mathbf{P}(A \cap B \,|\, \{x \not\leftrightarrow y\}) \;\leq\; \mathbf{P}(A \,|\, \{x \not\leftrightarrow y\})\mathbf{P}(B \,|\, \{x \not\leftrightarrow y\})\,. \tag{17}$$

Note the reversal of the inequality compared to Harris' inequality. The result seems intuitively plausible, since if we condition on $\{x \not\leftrightarrow y\}$, then further conditioning on the connected component C_x being large (in some sense) restricts the space available to C_y and should therefore tend to make it smaller. It appears, however, to be not such an easy challenge to find a more direct proof than the Markov chain argument given below. An alternative proof, based on induction in the size of the graph G, arises by concatenating vanden Berg and Kahn (2001), Proof of Thm. 1.2, and van den Berg et al. (2006a), Proof of Thm. 1.5.

In van den Berg et al. (2006a), Theorem 2 is proved in the greater generality of the random-cluster model with clustering parameter $q \geq 1$ (the case $q = 1$ corresponds to the ordinary bond percolation setup considered here) using an extension of the Markov chain argument; the induction-based alternative proof seems not to work in this setting. Applications of Theorem 2 to settle certain open problems concerning the equilibrium behavior of an interacting particle system known as the contact process appear in van den Berg et al. (2006a, 2006b).

Proof of Theorem 2: It suffices to prove the theorem for the case where G is finite, as the infinite case follows from standard limiting arguments similar to those discussed in the proof of Theorem 1. Consider the $\{0,1\}^E$-valued Markov chain $(X(0), X(1), \ldots)$ with the following transition mechanism, where to go from time k to time $k+1$, the edge configuration $X(k)$ is modified into $X(k+1)$ via an intermediate configurations $X'(k)$:

1. For each edge $e \in E$ that either is in the connected component of $X(k)$ containing x or has a vertex in this connected component as an endpoint, set $X'_e(k) = X_e(k)$, while for all other edges set $X'_e(k) = 1$ (resp. 0) with probability p (resp. $1 - p$), independently for different edges.

2. For each edge $e \in E$ that either is in the connected component of $X'(k)$ containing y or has a vertex in this connected component as an endpoint, set $X_e(k+1) = X'(k)$, while for all other edges set $X_e(k+1) = 1$ (resp. 0) with probability p (resp. $1-p$), independently for different edges.

Write μ for the probability measure on $\{0,1\}^n$ corresponding to conditioning percolation with parameter p on the event $\{x \not\leftrightarrow y\}$. If we condition μ on the connected component C_x contaning x, then clearly the conditional distribution of the rest of the configuration is i.i.d. (p) percolation on the edges that are neither in C_x, nor adjacent to a vertex in C_x. Viewing the above transition kernel as the composition of two kernels – Steps 1 and 2 above – we thus see immediately that μ is invariant under Step 1, and similarly under Step 2, and therefore also under the full kernel for $(X(0), X(1), \ldots)$. Furthermore, the chain is easily seen to be irreducible (within the set of states satisfying $x \not\leftrightarrow y$) and aperiodic, so no matter how the initial state $X(0)$ is chosen, we know that the distribution of $X(k)$ tends to μ as $k \to \infty$.

Now, if we were to imitate the approach of the previous two chapters, we would look for some suitable second Markov chain to couple $(X(0), X(1), \ldots)$ with. We will not do so here, but instead something similar in spirit, namely to specify in more detail how the randomization in the transition mechanism is carried out. To this end, we introduce an array $\{U_e(k)\}_{e\in E, k=0,1,2,\ldots}$ of i.i.d. Bernoulli (p) random variables, and an array $\{U_e^*(k)\}_{e\in E, k=0,1,2,\ldots}$ of i.i.d. Bernoulli $(1-p)$ random variables, independent also of the first array. (Coupling affectionados may view the following as a coupling of the Markov chain and these arrays.) The random parts of Steps 1 and 2 are implemented as follows:

1. For each edge $e \in E$ that is neither in the connected component of $X(k)$ containing x, nor incident to it, set $X'_e(k) = U_e(k)$.

2. For each edge $e \in E$ that is neither in the connected component of $X'(k)$ containing y, nor incident to it, set $X_e(k+1) = 1 - U_e^*(k)$.

Now start the Markov chain in some fixed state: for definiteness we take $X(0) \equiv 0$. With the chosen transition mechanism, the set of edges in the connected component of x in $X(1)$ becomes an increasing function of the variables $\{U_e(0)\}_{e\in E}$, while the set of edges in the connected component of y in $X(1)$ becomes a decreasing function of the variables $\{U_e(0)\}_{e\in E}$ and $\{U_e^*(0)\}_{e\in E}$. And proceeding by induction, we see that the set of edges in the connected component of x in $X(k)$ is an increasing function of the

variables $\{U_e(i)\}_{e\in E, i=0,\ldots,k-1}$ and $\{U_e^*(i)\}_{e\in E, i=0,\ldots,k-2}$, and that the set of edges in the connected component of y in $X(k)$ is a decreasing function of the variables $\{U_e(i)\}_{e\in E, i=0,\ldots,k-1}$ and $\{U_e^*(i)\}_{e\in E, i=0,\ldots,k-1}$.

Constructing the Markov chain as a (in parts) monotone function of i.i.d. random variables puts us in an ideal position for exploiting Harris' inequality. Write μ_k for the distribution on $\{0,1\}^E$ of $X(k)$, and fix two increasing events $A \in \mathcal{F}_x$ and $B \in \mathcal{F}_y$ as in the theorem. Noting that the composition of an increasing function and another increasing function is again increasing, we get from Harris' inequality (or from the slightly more elementary Proposition 1) that

$$\mu_k(A \cap B) \;\leq\; \mu_k(A)\mu_k(B)\,.$$

Sending $k \to \infty$ gives

$$\mu(A \cap B) \;\leq\; \mu(A)\mu(B)$$

and we are done. $\diamond$

Note that this proof shows a bit more than the statement of the theorem. If $f, g : \{0,1\}^E \to \mathbf{R}$ are bounded and increasing in the set of edges in the connected component containing x as well as decreasing in the set of edges that are in other connected components, then $\mu(fg) \geq \mu(f)\mu(g)$. If we restrict to the special case where f and g are functions only of the connected component containing x, then we recover an earlier result of van den Berg and Kahn (2001), Thm. 1.5.

Another aspect of the proof worth noting is that instead of the appeal at the beginning of the proof to "standard limiting arguments", we could alternatively have run the Markov chain directly on an infinite graph. The fact that μ_k converges in distribution to μ is then slightly less elementary than in the finite case, but still true. To see this, a classical-style coupling argument will do: Run the chain $(X(0), X(1), \ldots)$ starting from an arbitrary fixed configuration, together with another chain $(Y(0), Y(1), \ldots)$ with the same transition mechanism, with $Y(0)$ chosen at random according to μ. We may use the same Bernoulli variables $\{U_e(k)\}_{e\in E, k=0,1,2,\ldots}$ and $\{U_e^*(k)\}_{e\in E, k=0,1,2,\ldots}$ for the updating of the two chains. If x is incident to exactly d edges $e_1, \ldots, e_d$, then this guarantees that the two chains become identical from time $k+1$ and onwards as soon as $U_{e_1}(k) = \ldots = U_{e_d}(k) = 0$. This implies that for any event A that we have

$$|\mu_k(A) - \mu(A)| \;\leq\; \left(1 - (1-p)^d\right)^k\,, \tag{18}$$

which thus is a bound on the so-called total variation distance between μ and μ_k, and the key thing here is of course that this bound tends to 0 as $k \to \infty$.

5 Concluding remarks

The examples of Markov chain arguments in Sections 2–4 have a number of common features. One such feature is that each of them is used to derive one inequality or another. This raises the question of whether the circle of ideas that I try vaguely to define through these examples is limited to establishing inequalities. The answer is no, and the reader may, e.g., turn to the Proof of Lem. 2.4 in Häggström (1996) for an argument that can immmediately be recognized as belonging to the same circle, but is used to prove a distributional identity. Other examples of results of kinds other than inequalities obtained in similar fashion are the ergodic-theoretic results of van den Berg and Steif (1998), Häggström and Steif (2000), Häggström et al. (2000) and Häggström et al. (2002) mentioned at the end of Section 1.

Another noteworthy feature is the following. As in MCMC, some of the examples exploit the asymptotic behavior of their respective Markov chains. But unike in MCMC, where it is of crucial importance that the convergence to equilibrium is relatively fast, the rate of convergence to equilibrium is unimportant in each of the examples here. The difference is, of course, that in MCMC we need to implement and run the chains in computer simulations, while in the examples considered here we only need to run them "in our heads", where it only takes a split second to imagine having run them until we are close to equilibrium.

Acknowledgement. I am grateful to Yuval Peres and to Tue Tjur for pointing out the connection between Proposition 2 and Newton's inequality. Thanks also to Jeff Steif and to a referee for comments and corrections.

References

[1] Aires, N., Jonasson, J. and Nerman, O. (2002) Order sampling with prescribed inclusion probabilities, *Scand, J. Statist.* **29**, 183–187.

[2] Basharian, G.P., Langville, A.N. and Naumov, V.A. (2004) The life and work of A.A. Markov, *Linear Algebra Appl.* **386**, 3–26.

[3] van den Berg, J., Häggström, O. and Kahn, J. (2006a) Some conditional correlation inequalities for percolation and related processes, *Random Structures Algorithms* **29**, 417–435.

[4] van den Berg, J., Häggström, O. and Kahn, J. (2006b) Proof of a conjecture of N. Konno for the 1D contact process, in *Dynamics and Stochastics: Festschrift in Honor of Michael Keane* (Denteneer, D., den Hollander, F. and Verbitskiy, E., eds), IMS Lecture Notes–Monograph Series, pp 16–23.

[5] van den Berg, J. and Kahn, J. (2001) A correlation inequality for connection events in percolation, *Ann. Probab.* **29**, 123–126.

[6] van den Berg, J. and Kesten, H. (1985) Inequalities with applications to percolation and reliability, *J. Appl. Probab.* **22**, 556–569.

[7] van den Berg, J. and Steif, J.E. (1998) On the existence and non-existence of finitary codings for a class of random fields, *Ann. Probab.* **27**, 1501–1522.

[8] Esary, J.D., Proschan, F. and Walkup, D.W. (1967) Association of random variables, with applications, *Ann. Math. Statist.* **38**, 1466–1474.

[9] Fortuin, C.M., Kasteleyn, P.W. and Ginibre, J. (1971) Correlation inequalities on some partially ordered sets, *Comm. Math. Phys.* **22**, 89–103.

[10] Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Machine Intelligence* **6**, 721–741.

[11] Georgii, H.-O., Häggström, O. and Maes, C. (2001) The random geometry of equilibrium phases, *Phase Transitions and Critical Phenomena, Volume 18* (C. Domb and J.L. Lebowitz, eds), pp 1–142, Academic Press, London.

[12] Gilks, W., Richardson, S. and Spiegelhalter, D. (1996) *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

[13] Grimmett, G.R. (1999) *Percolation* (2nd edition), Springer, New York.

[14] Häggström, O. (1996) The random-cluster model on a homogeneous tree, *Probab. Th. Related Fields* **104**, 231–253.

[15] Häggström, O. (2002) *Finite Markov Chains and Algorithmic Applications*, Cambridge University Press.

[16] Häggström, O., Jonasson, J. and Lyons, R. (2002) Coupling and Bernoullicity in random-cluster and Potts models, *Bernoulli* **8**, 275–294.

[17] Häggström, O., Schonmann, R. and Steif, J.E. (2000) The Ising model on diluted graphs and strong amenability, *Ann. Probab.* **28**, 1111–1137.

[18] Häggström, O. and Steif, J.E. (2000) Propp–Wilson algorithms and finitary codings for high noise Markov random fields, *Combin. Probab. Comput.* **9**, 425–439.

[19] Harris, T.E. (1960) Lower bound for the critical probability in a certain percolation process, *Proc. Cambridge Phil. Soc.* **56**, 13–20.

[20] Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**, 97–109.

[21] Holley, R. (1974) Remarks on the FKG inequalities, *Comm. Math. Phys.* **36**, 227–231.

[22] Jonasson, J. and Nerman, O. (1996) On maximum entropy πps-sampling with fixed sample size, technical report, Chalmers and Göteborg University, `http://www.math.chalmers.se/Stat/Research/Preprints/index.cgi`

[23] Kallenberg, O. (1997) *Foundations of Modern Probability*, Springer, New York.

[24] Kesten, H. (1980) The critical probability of bond percolation on the square lattice equals $\frac{1}{2}$, *Comm. Math. Phys.* **74**, 41–59.

[25] Lindvall, T. (1992) *Lectures on the Coupling Method*, Wiley, New York.

[26] Markov, A.A. (1906) Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga, *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 2-ya seriya, tom 15, pp 135–156.

[27] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953) Equations of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087–1092.

[28] Newton, I. (1707) *Arithmetica Universalis: Sive de Compositione et Resolutione Arithmetica Liber.*

[29] Niculescu, C. (2000) A new look at Newton's inequalities, *J. Inequalities Pure Appl. Math.* **1**, Issue 1, Article 17.

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

[31] Reimer, D. (2000) Proof of the van den Berg–Kesten conjecture, *Combin. Probab. Comput.* **9**, 27–32.

[32] Thorisson, H. (1995) Coupling methods in probability theory, *Scand. J. Statist.* **22**, 159–182.

[33] Thorisson, H. (2000) *Coupling, Stationarity, and Regeneration*, Springer, New York.

Olle Häggström
Mathematical Sciences
Chalmers University of Technology
412 96 Göteborg
SWEDEN
`olleh@math.chalmers.se`