Stochastic cellular systems: ergodicity, memory, morphogenesis

edited by
R. L. Dobrushin, V. I. Kryukov and A. L. Toom

Manchester University Press
Manchester and New York

Distributed exclusively in the USA and Canada by St. Martin's Press

Published by Manchester University Press
Oxford Road, Manchester M13 9PL, UK
and Room 400, 175 Fifth Avenue,
New York, NY 10010, USA

Distributed exclusively in the USA and Canada
by St. Martin's Press, Inc.,
175 Fifth Avenue, New York, NY 10010, USA

British Library cataloguing in publication data
Stochastic cellular systems.
1. Organisms. Cells of analysis. Applications of stochastic processes
I. Dobrushin, R. L., *1929–* II. Kryukov, V. I. III. Toom, A. L. IV. Series
574.87'6

Library of Congress cataloging in publication data
Stochastic cellular systems: ergodicity, memory, morphogenesis/ edited by R. L. Dobrushin, V. I. Kryukov, and A. L. Toom.
p. cm. — (Nonlinear science)
Includes bibliographical references.
ISBN 0–7190–2206–1
1. Stochastic systems. 2. Markov processes. 3. Cellular automata. I. Dobrushin, R. L., 1929– . II. Kryukov, V. I., 1935– . III. Toom, A. L. IV. Series.
QA402.S845 1990
003'.76—dc20 89–12617

ISBN 0 7190 2206 1 *hardback*

Typeset in 10/12pt Times by
Graphicraft Typesetters Ltd., Hong Kong
Printed in Great Britain by
Biddles Ltd., Guildford and King's Lynn

Contents

Foreword

A new trend is under intense development in mathematics, both theoretical and applied. It relates to large systems consisting of many components which interact with each other locally. These systems may be deterministic or stochastic. From the probabilistic point of view, such systems are a special case of multidimensional Markov processes. Systems of this kind are designed to describe many forms of collective behaviour: physical, biological, and others.

Some specific processes of this kind were first considered in statistical physics, where they were used to model the dynamics of crystalline substances, their components being single spins of atoms. On the other hand, multicomponent systems appeared (perhaps, independently) in the course of attempts to develop models of biological processes, their components being formal neurons.

In the USSR, a powerful impetus to the study of such systems was given by I. M. Gelfand and especially by his student M. L. Tsetlin whose reflections on the nature and problems of collective behaviour influenced all of us. In fact, we should speak of a 'Tsetlin movement' which existed for many years, and our work is much indebted to this movement. Also, about two decades ago it became clear to mathematicians that multicomponent stochastic systems with local interaction should be investigated as abstract mathematical entities without taking into account any applied interpretation. This does not prevent one from thinking of them in the context of applications; on the contrary, abstract examination seems to have promoted many new applications. Indeed, every year we read about some new topics where multicomponent systems have become useful. Methods of multicomponent stochastic systems theory are being applied in many parts of biology, on the molecular, cellular, genetic and

population levels; in computer science; in polymer chemistry; in seismology; in economics, and other sciences. Similar mathematical constructions are now widely investigated under various names, including homogeneous random media, cellular automata, tessellations, and honeycomb structures.

It seems that all or many of the studies related to our topic have common features. Let us formulate our three main principles, which define the scope of the present volume: all systems treated here are multicomponent (that is, consist of many or infinitely many components), homogeneous (all components have similar rules of behaviour), and local (every component's behaviour is a random function of its nearest neighbours' states only).

The problems we try to solve have much in common too. From the models of statistical physics we have inherited interest in phase transitions, that is, discontinuous changes of the system's parameters as a result of continuously changing parameters of a single component. The phenomenon of metastability is typical of our systems too: this in particular means that a system behaves for a long time as if it had reached the proper equilibrium state, but then the system jumps to another quasi-ergodic mode of behaviour.

Bearing in mind the importance of developing contacts between mathematicians and biologists, of mutual enrichment in ideas and promoting biologically oriented mathematical investigations, we would mention the seminars on the theory of multicomponent random systems and its applications to biology which are regularly organized in the Centre of Biological Researches in Pushchino by two Academic Institutes: the Research Computing Centre (Pushchino) and the Institute for Problems of Information Transmission (Moscow). The first seminar was held in 1976, and its works were published as *Vzaimodeistvuyushchie Markovskie prozessy v biologii*, Pushchino, 1977 (in Russian) and later as volume 653 in the Springer series 'Lecture Notes in Mathematics' under the title *Locally Interacting Systems and their Application in Biology*, 1978. Further seminars were held in 1978, 1981 and 1984. Their publications were special volumes in Russian, some of which were later translated into English. Recently A. V. Holden of the Centre for Nonlinear Studies at Leeds University has offered us the publication of the present volume to present the main developments of our seminars in the vein of the previous Springer work.

This volume is based on lectures and reports of all the four seminars. Regrettably we failed in our initial project of presenting all the ideas, methods and applications in a coherent form in a single monograph. It seems difficult to achieve at the present stage of our work, when the parts are still separated and one can only dream of compiling them into a harmonious whole. Instead, each group of authors describe their work in

detail, hoping that the sensible reader will hit on the general principles that unite us.

The mathematical theory of interacting Markov processes, which is closely associated with the general theory of Gibbs random fields, seems to be the most elaborate of our developments. These processes may have discrete or continuous time coordinate. Of course, these two cases have much in common, and the general framework, which subsumes both, has been developed both in the USSR and abroad; but every case has its own peculiarities. Further, the continuous-time case has been elaborated in Western studies and is well covered by a number of surveys and monographs, the latest of which is the well-thought-out *Interacting Particle Systems*, Springer, 1985 by T. M. Liggett. Many (perhaps, most) results on the discrete-time interacting Markov processes belong to Soviet workers. This line of thought emerged earlier than the continuous-time one, but for a long time remained unsummed up and insufficiently known abroad; the present volume's first part is the survey of the theory of discrete-time interactive random processes. The authors have tried to condense the bulk of the work into a few chapters. In this way Part I contains four major sections on proofs of ergodicity, proofs of non-ergodicity, Bernoulli and Markov invariant measures, and algorithmic methods.

The second part is the first survey on the topic which is a bridge between multicomponent random systems and the theory of queues. The components here are processors which receive and exchange requests to be served, while unserved requests form queues. Here the field of application is quite different, but the mathematical apparatus is nevertheless analogous to that developed in the other parts of the volume.

The third part is devoted to the application of the theory of locally interacting Markov processes to the modelling of collective effects in the brain. Similar development is well known in modern computational neurobiology as the Hopfield (spin glass) paradigm. But there are some essential differences, other than geographical or historical particularities. First, the commonly accepted assumption that the nerve cell can be described by a formal neuron analogue is too narrow for us to account for a large body of neurophysiological data. Secondly, the phase transitions and corresponding loss of stability in neuronal nets is here considered as a quite usual, though very important part of the nervous system mechanism, rather than as a source of catastrophe. Thirdly, the main principle of both local and global integration in neuronal nets is assumed to be not the famous principle of hierarchy but the metastability principle discovered by the Russian physiologist A. A. Ukhtomskii and known as the principle of the dominant.

The fourth and last part of our volume pertains to mathematical modelling of morphogenesis; this work also has not yet been surveyed

systematically. The idea of modelling morphogenesis is popular among mathematicians: recall the works of R. Thom and C. Zeeman among others. But the present approach is quite different, as it is based on the idea of local homogeneous interaction, and this descends from the Tsetlin movement. Our models of morphogenesis take advantage of the hypothesis (which is confirmed by a lot of data) that the actual biological morphogenesis often depends on some local homogeneous interactions between cells; therefore in our models every cell moves depending on a few neighbour cells (locality) and this dependence is defined by one and the same rule for all (or almost all) cells (homogeneity). Unlike systems in the other parts of this volume, the present systems are deterministic; but they conform to our main principles: they are multicomponent, homogeneous and local. Homogeneity is their key property, as it enables us to apply representation theory, which is the main mathematical tool in this work. Part IV treats circle and sphere formation as mathematical models of an important biological process – blastula formation. These models have led to two new interesting structures on graphs. One structure, which we call the 'graph with atlas', is induced by the rules according to which the graph's vertices move. The other structure, called the 'graph with rotation', is connected with the presentation of a biological tissue in the shape of a two-dimensional net, which lies on a surface in three-dimensional space $\mathbb{R}^3$; this structure is induced by the orientation of this surface. We found it interesting to learn that graphs with rotation had been previously applied in an elegant manner to the map colouring problem in the works of Ringel and Youngs; however, our definition of graphs with rotation is more general than theirs.

I. I. Piatetski-Shapiro has considerably influenced the first and last parts of this volume. Many of us are greatly indebted to him.

We thank the Pushchino Research Computing Centre, the Institute for Problems of Information Transmission and the Belozerski Laboratory of Moscow State University for supporting our work.

Contributors

Part I

A. L. Toom, N. B. Vasilyev and O. N. Stavskaya
Moscow State University, Belozerski Laboratory, Moscow GSP, 119899, USSR

L. G. Mityushin, G. L. Kurdyumov and S. A. Pirogov
Institute for Problems of Information Transmission of the USSR Academy of Science, 19 Ermolova Street, Moscow, GSP-4, 101447, USSR

Part II

R. L. Dobrushin, A. N. Rybko and Yu. M. Suhov
Institute for Problems of Information Transmission of the USSR Academy of Science, 19 Ermolova Street, Moscow GSP-4, 101447, USSR

M. Ya. Kelbert
International Institute of Earthquake Prediction Theory and Mathematical Geophysics, USSR Academy of Science, 79, k.2, Warshavskoye sh., Moscow 113556, USSR

Part III

V. I. Kryukov, G. N. Borisyuk, R. M. Borisyuk, A. B. Kirillov and Ye. I. Kovalenko
Research Computer Center, USSR Academy of Sciences, Pushchino, Moscow Region 142292, USSR

Part IV

A. M. Leontovich
Moscow State University, Belozerski Laboratory, Moscow GSP, 119899, USSR
V. B. Philippov
Central Institute for Research in Complex Automatization, Moscow, USSR

Part I
Discrete local Markov systems

A. L. Toom, N. B. Vasilyev, O. N. Stavskaya, L. G. Mityushin, G. L. Kurdyumov, S. A. Pirogov

I.1
Introduction

This survey is based on a vast cluster of papers by Soviet workers on multicomponent random systems. Objects of this kind are widely investigated in the USSR and many other countries, but we mention only those papers which are most relevant to the problems we concentrate on here. The results are grouped thematically to make our survey coherent; from the Introduction the reader may skip to Chapter 3, 8, 14, or 16.

Our main concern is the case when space and time are discrete, the set of states of a single site (automaton) is finite, and their interdependence during a unit of time is finite; we make only a few remarks about extensions to continuous time. The pioneer work for our school was [78].

We are deeply grateful to all who helped us to work on this survey, of whom A. M. Leontovitch should be singled out for special mention.

Chapter 1
Examples and ideas

In this introductory chapter we describe in informal terms some examples of 'automata', our interest in which guided the work, and show using these examples what kind of results were obtained.

Mainly we treat the homogeneous case where all automata in a system are identical and have a finite set of states. Unless otherwise specified, assume that every automaton has two states: 0 and 1. The state of the whole system is defined if states of all the automata are defined. Automata are placed at the vertices of an integer lattice (say, one- or two-dimensional) and are enumerated by one index $i \in \mathbb{Z}$ or by two indices $(i,j) \in \mathbb{Z}^2$. The time is discrete. For simplicity, we first explain the functioning of a homogeneous automata system, placed at the vertices of a one-dimensional lattice.

Every automaton having index $i \in \mathbb{Z}$ has a finite set of neighbours indexed by

$$i + u_1, i + u_2, \ldots, i + u_R.$$

(The automaton may or may not be its own neighbour.) The next moment of time $t + 1$ the ith automaton is in the state 1 with probability θ which is a function of the states of its neighbours at time t:

$$\theta = \theta(x^t_{i+u_1}, x^t_{i+u_2}, \ldots, x^t_{i+u_R}).$$

The automaton is in state 0 with the complementary probability $1 - \theta$. Let us add the basic condition that for any given states of all the automata at time t the states of all the automata at time $t + 1$ are chosen independently of each other. We term θ the transition function, its values are transition probabilities. It is important that the set of automata is infinite or very large, while the number R of every automaton's neigh-

bours is restricted; this is what is meant by locality of interaction.

Behaviour of such a system can be described in terms of the P operator which transforms the probability distribution at time t into the probability distribution at time $t + 1$. Due to the independence condition, our function θ defines P uniquely. If all the transition probabilities are equal to 0 or 1, we have a system of deterministic automata. In this case P is a map of the system into itself. If some values of θ equal 0 or 1, we call P degenerate, otherwise P is non-degenerate.

A general problem about systems of random automata consists in predicting their behaviour a long time after their functioning starts. The roughest, qualitative aspect of this problem is the ergodicity problem of such systems, which is our main concern here. We term a system ergodic if it 'forgets' its initial state, i.e. in the limit $t \to \infty$ tends to some probability distribution, which does not depend on the initial state. It can be proved (it follows from the fixed point theorem, in particular) that every one of our automata systems has at least one invariant measure, that is, such a probability distribution which turns into itself during one unit of time. Any ergodic system has just one invariant measure. Most results surveyed here are about ergodicity and invariant measures. More concretely, the survey presents:

A. General methods to prove ergodicity, which are applicable if the θ function depends on its arguments sufficiently weakly in one sense or another. Two ideas of such proofs will be explained in the application to Examples 1.1a and 1.3.
B. Combinatorial-geometrical proofs of non-ergodicity, which are similar to the contour method in lattice models of statistical physics. An idea of such a proof will be applied to Example 1.2.
C. Algorithmic methods. Even for a rather narrow class of systems the problem of recognising ergodic ones among them has been proved to be unsolvable. The proof includes simulation of Turing machines by such systems.

In some cases it is possible to present explicit description of invariant measures as they are Markov (in the one-dimension case) and Gibbs (in the two-dimension case) invariant measures. Some of these cases are presented as Examples 1.3 and 1.4. The Example 1.4b is tied up with the two-dimensional Ising model; its system is non-ergodic as it has at least two invariant measures.

Now let us pass to our examples. Remember, that all of them treat independent (termed synchronous in [7]) transitional operators; that is, all automata take their states independently if their last time states are fixed.

Example 1.1 ('Voting') [58]. Let every automaton have an odd number of neighbours. At every unit of time the following two transformations are performed.

First, every automaton assumes that state (0 or 1) in which the majority of its neighbours has just been. Second, every automaton changes its state (0 to 1, or 1 to 0) with probability ε independently from all the other automata and all the past events. On can express it by saying that all automata are unreliable and make random independent errors.

Our reasons for examining voting systems were the following. We did not know whether non-degenerate non-ergodic systems exist and believed it would be interesting to find some. It seemed that the 'voting' rule with small ε should be the best remedy against forgetting the initial state. (Chapter 13 will show that reality is more complicated.)

With ε near enough to $\frac{1}{2}$ all voting systems are ergodic. So our question is: Does such an ε, $0 < \varepsilon < 1$, exist, with which some voting system is non-ergodic? Computer simulation shows that this depends on the system's dimension. In fact the following three cases were simulated.

(a) Automata occupy integer points of a line. The automaton $i \in \mathbb{Z}$ has three neighbours: $i - 1$, i, $i + 1$. In this case simulation has showed ergodicity for all ε, $0 < \varepsilon < 1$.

In the two other cases automata are placed at the points of an integer plane and the automaton $(i,j) \in \mathbb{Z}^2$ has the following neighbours:

(b) three automata (i,j), $(i-1,j)$, $(i,j-1)$;

(c) five automata (i,j), $(i-1,j)$, $(i,j-1)$, $(i+1,j)$, $(i,j+1)$.

Simulation showed the cases (b) and (c) to be non-ergodic for sufficiently small ε. This has been proved, however, for the case (b), but not for (c). On the other hand, ergodicity has been proved:

for the case (a) with $\frac{1}{4} < \varepsilon < \frac{3}{4}$;

for the case (b) with $\frac{1}{3} < \varepsilon < \frac{2}{3}$;

for the case (c) with $\frac{2}{5} < \varepsilon < \frac{3}{5}$.

In fact finite automata systems were simulated by the Monte Carlo method. Say, we had the initial state 'all zero', some scores of automata and some thousands units of time; if the proportion of zeros remained clearly more than $\frac{1}{2}$ in this process, we inferred non-ergodicity; if the proportion of zeros approached $\frac{1}{2}$ quickly we inferred ergodicity.

As a matter of fact, a finite voting system is ergodic with all ε, $0 < \varepsilon < 1$. Therefore, interpreting the finite systems simulation data, one can speak of the velocity of their convergence to the limit distribution and of its dependence on the boundary conditions only, the value of ε, the

number of automata and the boundary conditions playing the role of parameters. It is natural to conjecture that finite systems converge quickly and their limit distributions do not depend on the boundary conditions for the same values of ε for which the infinite system is ergodic. In this survey we shall neither discuss this hypothesis nor even formulate it precisely. We note only that it is confirmed by all the cases known to us where ergodicity or non-ergodicity has been proved.

Let us say something about one way to prove ergodicity [96] for the case (a) of Example 1.1. Assume that we have two copies of our automata systems. Both systems start working at the same time, but one has the initial state 'all zeros', and the other 'all ones'. The possible mutual approaching of the two systems for $t \to \infty$ warrants ergodicity because all other initial states are 'between' our two.

Assume $\varepsilon \leqslant \frac{1}{2}$ (the case $\varepsilon > \frac{1}{2}$ can be reduced to this one). In this case the 'errors' of the two systems can be made mutually dependent (errors inside each system remaining independent) in such a way that for all realisations the condition $x_i^t \leqslant y_i^t$ will hold for all i and all t, where x_i^t and y_i^t are states of the ith automaton at time t in the first and second system respectively. This does not prevent every single system from working as before. This is done by introducing mutually independent random variables ω_i^t, each of which takes three values:

0 with probability ε,
$\frac{1}{2}$ with probability $1 - 2\varepsilon$,
1 with probability ε.

These variables control errors of both systems in the following way:

if $\omega_i^t = \frac{1}{2}$ both automata x_i and y_i make no errors at time t;
if $\omega_i^t \neq \frac{1}{2}$ both automata x_i and y_i go to the state equal to ω_i^t at time t (so they make errors if their states after voting were different).

Under these rules our two systems can be seen as one system of 'double automata' having actually just three states: 00, 01, 11. The proportion of 01 states in this double system tends to 0 as $t \to \infty$ if and only if the system in the example (a) is ergodic. Now note that the states 01 multiply no more than three times in a unit of time. Hence, if they 'die out' (due to errors) with rate greater than $\frac{2}{3}$, their proportion tends to 0. This is guaranteed if $\frac{1}{3} < \varepsilon < \frac{2}{3}$. A more complicated argument proves ergodicity with $\frac{1}{4} < \varepsilon < \frac{3}{4}$. But our conjecture of ergodicity for all ε, $0 < \varepsilon < 1$ is not yet proved.

We term a system non-degenerate if all its transitional probabilities differ from 0 and 1, and it is degenerate otherwise. The system in Example 1.1a is non-degenerate with $0 < \varepsilon < 1$. Thus, according to our

simulation data, this system is ergodic if and only if it is non-degenerate. For a very similar system (described below as Example 1.4a ergodicity for arbitrary small $\varepsilon > 0$ can be proved.

Of course, degenerate one-dimensional systems may be non-ergodic (like degenerate Markov chains). But it is interesting that the transition from ergodicity to non-ergodicity may occur at some intermediate (that is, different from 0 and 1) value of the parameter (which is impossible for finite Markov chains). This is illustrated by the following example, which we have studied a great deal.

Example 1.2 (Stavskaya's problem) [78]. Automata are placed in a line $\mathbb{Z}$. Neighbours of the ith automaton are $i - 1$ and i. The transition from t to $t + 1$ consists of two transformations.

First, for all $i \in \mathbb{Z}$, the automaton i takes the state equal to $x_{i-1}^t \cdot x_i^t$ (which equals 1 only if $x_{i-1}^t = x_i^t = 1$, and equals 0 otherwise). Second, every automaton, having got into the state 0, makes an 'error' that changes its state into 1 with probability ε independently from all the other automata and all the past events. Automata, having got into the state 1, never make errors.

According to what has been said, this system is ergodic, when ε is near 1. But it is interesting that when ε is near 0, the system is non-ergodic. Let us describe its limit behaviour in more detail. Evidently the state **1** (all automata are ones) goes into itself with the probability 1, whence the measure $\delta_{\mathbf{1}}$ concentrated in this state is invariant. Denote by μ_ε the limit $t \to \infty$ measure if at $t = 0$ the system was in the state **0** ('all automata are zeros') – for this particular system existence of this limit can be proved easily. Denote by $\mu_\varepsilon(1)$ the probability (which does not depend on i) that $x_i = 1$ in μ_ε. It has been proved [71,85] that there is such a 'critical value' ε^* of ε, which seems to be about 0.31 (it has been proved that $0.09 < \varepsilon^* < 0.323$ [85]), for which the following assertions hold:

(1) With $\varepsilon > \varepsilon^*$ our system is ergodic. In this case $\mu_\varepsilon(1) = 1$, that is, $\mu_\varepsilon = \delta_{\mathbf{1}}$.
(2) With $\varepsilon < \varepsilon^*$ our system is non-ergodic. In this case $\mu_\varepsilon(1) < 1$, that is, μ_ε differs from $\delta_{\mathbf{1}}$. All homogeneous (that is, translation-invariant) invariant measures are of the form $\alpha\mu_\varepsilon + (1 - \alpha)\delta_{\mathbf{1}}$ where $0 < \alpha < 1$ [47].

It is plausible that all invariant measures of this system are homogeneous and that μ_ε depends on ε analytically in the range $0 < \varepsilon < \varepsilon^*$, but both assertions are proved only for sufficiently small ε [100]. The measure μ_ε has been proved to depend on ε continuously in $\varepsilon < \varepsilon^*$ [47], but it is unknown whether this dependence is continuous at $\varepsilon = \varepsilon^*$. The measure μ_ε for $\varepsilon < \varepsilon^*$ is neither a Markov one, nor a k-Markov one for any k [84].

Let us sketch a proof that the Example 1.2 system is non-ergodic with some small $\varepsilon > 0$ [85]. Let us start from the state 'all zeros'. Assume that every pair (i,t), $i \in \mathbb{Z}$, $t \in \mathbb{Z}_+$ has its own variable $\omega_i^t \in \{0;1\}$ (see Fig. 1.1, where crosses mark points (i,t) where $\omega_i^t = 1$). Now these variables will be the only source of randomness in the system. They control automata behaviour in the following way:

if $\omega_i^t = 0$, the automaton i makes no error at time t;
if $\omega_i^t = 1$, the automaton i is in the state 1 at time t, even if it has to make an error to do so.

Thus, the state of the automaton 0 at time T is some function of a finite number of variables ω_i^t, namely of those in which $0 < t \leqslant T$, $T - t \leqslant i \leqslant 0$. Our main point is the following representation of this function: the automaton 0 at time T is in the state 0, that is, $x_0^T = 0$, if and only if there is a path from the point $(0,T)$ to the line $t = 0$, consisting of two kinds of steps only: down, at vector $(0;-1)$ or left-down, at vector $(-1;-1)$, along which all points have the corresponding variable $\omega_i^t = 0$. In other words, we imagine that from every point (i,t) of our integer lattice two pipes start: one goes down to the point $(i,t-1)$, the other goes left-down to the point $(i-1,t-1)$, and there are stoppers at those points where $\omega_i^t = 1$. Then the automaton 0 at time T will be in state 0 if and only if it is possible for a fluid to percolate from the point $(0,T)$ to the line $t = 0$. The automaton 0 at time T will be in state 1 if the point $(0,T)$ is fenced off from the line $t = 0$, which implies that stoppers form a 'fence'.

Geometrically, this fence is placed along a polygonal line consisting of three types of links, which are directed:

(a) left-down, at vector $(-1;-1)$;
(b) right, at vector $(1;0)$;
(c) up, at vector $(0;1)$.

We may assume that every one of these polygonal lines begins and ends at the point $(\frac{1}{2},T+1)$. The centre of every type-(b) link is a point (i,t), where $\omega_i^t = 1$, that is, there is a stopper. (See example of a fence in Fig. 1.1.)

Now estimate the probability that $x_0^T = 1$. It does not exceed the sum of probabilities of all possible fences, that is $\sum_{k=1}^{\infty} N_k \varepsilon^k$ where N_k is the number of fences having k stoppers. If a fence has k stoppers, its polygonal line has k type-(b) links; hence it has just k type-(a) links and k type-(c) links. Thus, $N_k \leqslant 3^{3k}$. So, with $\varepsilon < 1/54$, our sum $\sum_{k=1}^{\infty} N_k \varepsilon^k$ is

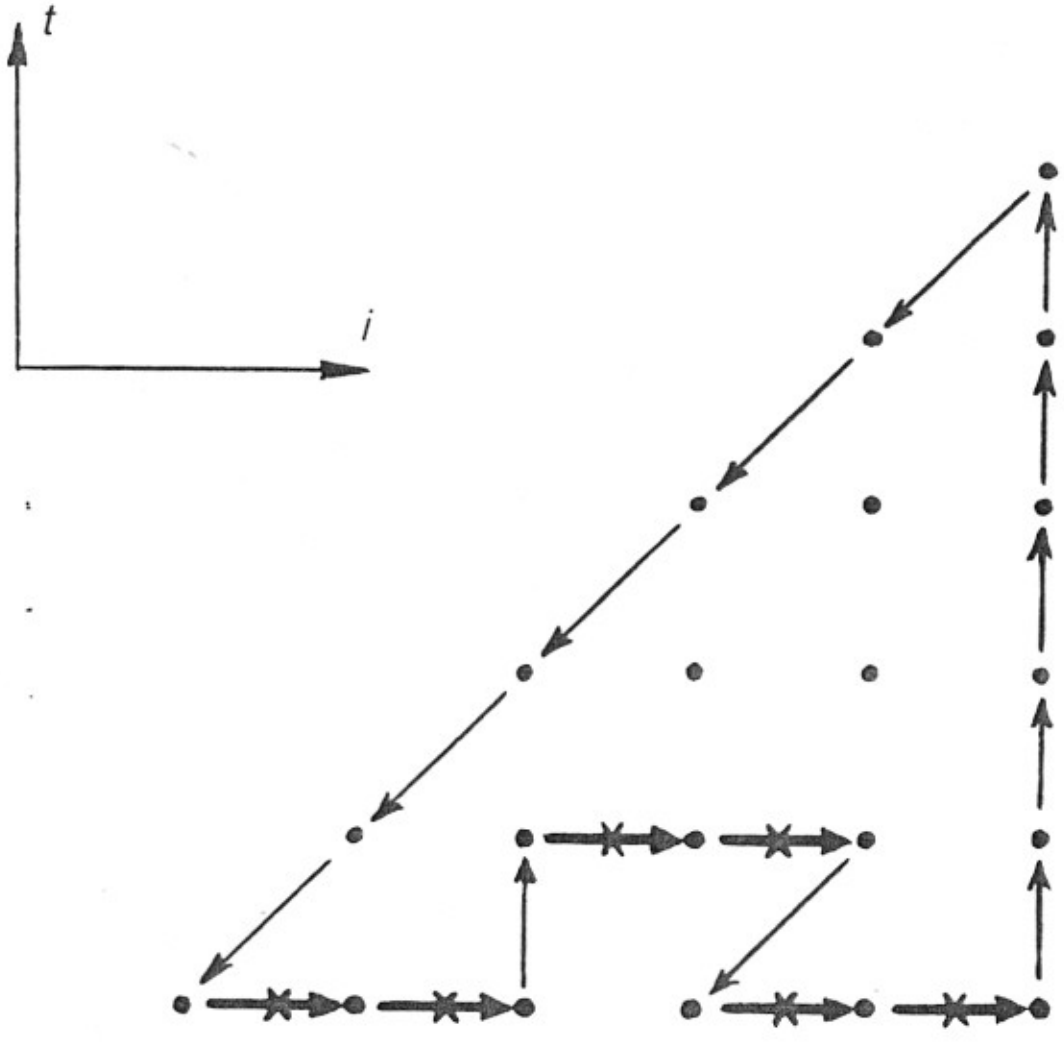

Fig. 1.1 A 'fence'. See Example 1.2.

less than 1, which guarantees that the proportion of zeros in our automata system does not tend to 0 and the system is non-ergodic.

This sketch shows that investigation of our systems has connections with percolation theory problems (see, for example, [41]).

In fact, investigation of the Example 1.2 system, among others, began with its computer simulation, which showed non-ergodicity with small values of ε and ergodicity with large enough values of ε. Thus, here too we may hypothesise that finite systems converge quickly if and only if the infinite system is ergodic.

The finite systems analogous to the Example 1.2 system have been proved to converge quickly with large ε and slowly with small ε. More exactly, denote by $\tau_\varepsilon(n)$ the mean time when the system of n automata having started from the state 'all zeros' first gets into the state 'all ones'. Then

$$c_1 \ln n < \tau_\varepsilon(n) < c_2 \ln n + c_3 \quad \text{for large enough } \varepsilon \text{ [78]};$$
$$\exp(c_4 n) < \tau_\varepsilon(n) < \exp(c_5 n) \quad \text{for small enough } \varepsilon \text{ [85]}.$$

Here c_1, c_2, c_3, c_4, c_5 are positive parameters depending only on ε.

In Examples 1.3 and 1.4 which follow, the investigation of ergodicity is

promoted by the possibility of describing an invariant measure explicitly: in Example 1.3 it is Bernoulli, in Example 1.4a it is 2-Markov, in the two-dimensional Example 1.4b it is Gibbs.

Example 1.3 [96]. As in the preceding example, every automaton $i \in \mathbb{Z}$ has two neighbours $i - 1$ and i, but the transformation from t to $t + 1$ is different.

First, every automaton i goes to 1 if $x_{i-1}^t \neq x_i^t$ and goes to 0 if $x_{i-1}^t = x_i^t$. Thus, it takes the state $x_{i-1}^t \oplus x_i^t$ where $\oplus$ means sum modulo 2. Second, every automaton changes its state (makes an error) with probability ε independently from all other automata and from all past events.

This system has been proved to be ergodic with all ε, $0 < \varepsilon < 1$, and its only limit measure is easily described: it is Bernoulli, that is, automata states are independent random variables, equal to 0 and 1 with equal probabilities $\frac{1}{2}$. Let us prove it. Introduce the following functions of the system state

$$x = (x_i),\ i \in \mathbb{Z}\colon\ \chi_i(x) = (-1)^{x_i},\ \chi_K = \prod_{i \in K} \chi_i$$

where K is any finite subset of $\mathbb{Z}$.

It is sufficient for us to prove for any K that the mean value $\mathbf{E}\chi_K$ of χ_K at time t tends to 0 when $t \to \infty$ (for any initial state of the system). Denoting the system states at times $t - 1$ and t by x^{t-1} and x^t we write

$$\mathbf{E}\chi_i(x^t) = (1 - 2\varepsilon)\chi_i(x^{t-1})\chi_{i-1}(x^{t-1}), \tag{1.1}$$

$$\mathbf{E}\chi_K(x^t) = \prod_{i \in K} \mathbf{E}\chi_i(x^t). \tag{1.2}$$

The equation (1.1) holds because $\chi_i(x^t)$ equals $\chi_i\chi_{i-1}(x^{t-1})$ with probability $1 - \varepsilon$, hence equals $-\chi_i\chi_{i-1}(x^{t-1})$ with probability ε. The equation (1.2) follows from the independence of the automata choices. Now, (1.1) and (1.2) give us

$$\mathbf{E}\chi_K(x^t) = (1 - 2\varepsilon)^{|K|}\chi_{K^1}(x^{t-1}) \tag{1.3}$$

where K^1 consists of those j for which just one of the two numbers j, $j + 1$ belongs to K; $|K|$ is the cardinality of K.

Double iteration of (1.3) gives

$$\mathbf{E}\chi_K(x^t) = (1 - 2\varepsilon)^{|K|}\mathbf{E}\chi_{K^1}(x^{t-1}) = (1 - 2\varepsilon)^{|K|+|K^1|}\chi_{K^2}(x^{t-2})$$

where K^2 is obtained from K^1 by the same rule by which K^1 was obtained from K. We can iterate (1.3) further. Thus we obtain

$$\mathbf{E}\chi_K(x^t) = (1 - 2\varepsilon)^{|K|+|K^1|+\ldots+|K^t|}\chi_{K^t}(x^0)$$

where every K^{s+1} is obtained from K^s by the same rule. Evidently, $K \neq \varnothing$ implies $|K^1| \geqslant 2$, whence $|K^s| \geqslant 2$ for all $s \geqslant 1$. Thus, $|\mathbf{E}\chi_K(x^t)| \leqslant (1 - 2\varepsilon)^{2t+1}$ which tends to 0 with $t \to \infty$.

Consider the value of

$$\sigma_t = |K| + |K^1| + \ldots + |K^{t-1}|.$$

In the special case $|K| = 1$ it is the number of ones in the first t lines of the Pascal triangle modulo 2, and for all non-empty K it grows in the same way:

$$\sigma_t > \text{const} \cdot t^{\log_2 3}$$

This provides super-exponential convergence to the limit Bernoulli measure in our system: probabilities of cylinder sets differ from their limit values less than

$$\text{const} \cdot (1 - 2\varepsilon)^{t^{\log_2 3}}$$

This (function) method can be extended to many other systems to prove ergodicity by choosing appropriate base functions χ_k and using our iteration argument.

Note another property of transitional probabilities in Example 1.3. Let the state x_{i-1}^t be fixed one way or the other: $x_{i-1}^t = 0$ or $x_{i-1}^t = 1$, while x_i^t take the values of 0 and 1 with equal probabilities $\frac{1}{2}$; in this case x_i^{t+1} takes the values of 0 and 1 with equal probabilities $\frac{1}{2}$ also. This property can be generalised in the following way to construct a class of ergodic systems having Bernoulli limit measures [97].

Let every automaton $i \in \mathbb{Z}$ have two neighbours $i - 1$ and i and n states $\{1,2,\ldots,n\}$. Let such a distribution ξ on this state set exist that with any fixed value of x_{i-1}^t, the value of x_i^t being thus distributed implies the same distribution ξ of the value of x_i^{t+1}. In these assumptions, the system is provably ergodic and its limit measure is a Bernoulli one, that is, a product of measures identical with ξ. Of course, the same holds when x_{i-1} being distributed in ξ implies x_i^{t+1} being distributed in the same way with any fixed x_i^t.

Consider all homogeneous one-dimensional systems in which every automaton $i \in \mathbb{Z}$ has $i - 1$ and i as neighbours, and n states. This family of systems has $n^3 - n^2$ parameters:

$$\theta(s|\ell,r) = \mathbb{P}(x_i^{t+1}=s|x_{i-1}^t=\ell,x_i^t=r),$$

where $1 \leqslant \ell, r \leqslant n$, $1 \leqslant s \leqslant n - 1$. We have described the subset of this family consisting of systems having Bernoulli or Markov invariant measure. This is not trivial because the direct verification whether a given measure is invariant for a given system consists in checking an infinite system of equalities. But in fact, for a Markov measure, this verification provably boils down to checking some finite system of equalities corresponding to the behaviour of a string of $n + 1$ automata. The set of combined values of $\theta(s|\ell,r)$ for which the system has Markov invariant measure is a piece of some algebraic manifold of dimension $(n - 1)(n^2 - n + 1)$. For all $n > 2$ the two components of this manifold

correspond to systems having Bernoulli invariant measures of the kind described above (reproducing the distribution of one neighbour with any fixed value of the other neighbour). Note that ergodicity of these systems having Bernoulli invariant measures has been proved (and has also been done for some analogous d-dimensional systems too), but the ergodicity question remains open for other systems with Markov invariant measures.

The special case $n = 2$ of two-neighbour systems having Bernoulli and Markov invariant measures was examined first in [2]. With $n = 2$ a system is defined by four parameters:

$$\theta(1|0,0),\ \theta(1|0,1),\ \theta(1|1,0),\ \theta(1|1,1).$$

Their combined values can be represented as points in a four-dimensional cube. In this cube there are two non-reducible manifolds of dimension 3 (in which systems having Bernoulli invariant measures can be found) and another manifold also of dimension 3 (in which systems having Markov invariant measures can be found).

Example 1.4 [42]. This example, in which every automaton has two states 0 and 1 as before, will remind us of the voting in Example 1.1. Consider first the one-dimensional version.

(a) Every automaton $i \in \mathbb{Z}$ has three neighbours $i - 1$, i, $i + 1$ and its probability to be in the state 1 at time $t + 1$ is a function θ of the number r of its neighbours having been in the state 1 at time t, as presented in Table 1.1. (The bottom line gives the corresponding values of Example 1.1a, for comparision.) This system has been proved to be ergodic with all ε, $0 < \varepsilon \leqslant 1$. The case $\varepsilon = 1$ is evident, the cases with small ε are our concern; in these cases Examples 1.1a and 1.4a are especially similar. The proof of its ergodicity is based on the following presentation of its invariant measure. It is more convenient to describe not just the measure, but the corresponding joint distribution for times t and $t + 1$. It has the following property: for any natural m and any states

$$x^t_{-m},\ x^{t+1}_{-m},\ x^t_m,\ x^{t+1}_m$$

the conditional probability of certain values

$$(x^t_i),\ (x^{t+1}_j),\ -m + 1 \leqslant i, j \leqslant m - 1$$

Table 1.1

r	3	2	1	0
θ	$\dfrac{1}{1+\varepsilon^3}$	$\dfrac{1}{1+\varepsilon}$	$\dfrac{\varepsilon}{1+\varepsilon}$	$\dfrac{\varepsilon^3}{1+\varepsilon^3}$
θ	$1-\varepsilon$	$1-\varepsilon$	ε	ε

is proportional to the product of values

$$\alpha(x_i^t, x_j^{t+1}) = \varepsilon^{(x_i^t + x_j^{t+1} - 1)/2}$$

over all such pairs (i,j) that $-m \leqslant i \leqslant j \leqslant m$ and $|i - j| \leqslant 1$. (In Fig. 1.2 these pairs are connected with lines.) In physical terms it is a Gibbs distribution with pair potential

$$\varphi(x_j^t, x_j^{t+1}) = \ln \alpha(x_i^t, x_j^{t+1}).$$

Now let us pass to two-dimensional systems. As we have said, some non-degenerate two-dimensional systems are non-ergodic. The five-neighbour voting Example 1.1c is not proved to be non-ergodic, although it seems to be. But the following similar system has been proved to be non-ergodic:

(b) Every automaton $(i,j) \in \mathbb{Z}^2$ has four neighbours:

$$(i,j),\ (i-1,j),\ (i,j-1),\ (i-1,j-1).$$

At time $t + 1$ the automaton (i,j) gets into the state 1 with probability θ which depends on the number r of its neighbours having been in the state 1 at time t as in Table 1.2.

Fortunately, some Gibbs measure fits as the joint distribution at times t and $t + 1$ of this system. This Gibbs measure corresponds to some two-dimensional Ising model with attraction, and our parameter ε is proportional to $\exp(2\beta)$ where β^{-1} is the model temperature. Vertices of the two square two-dimensional lattices corresponding to times t and $t + 1$ may be thought of as white and black squares of an infinite chessboard on which the Gibbs measure is defined. It is well known that temperature β^{-1} being small (less than 'phase transition' point) the Ising model has two different Gibbs measures. Hence our automata system with ε being

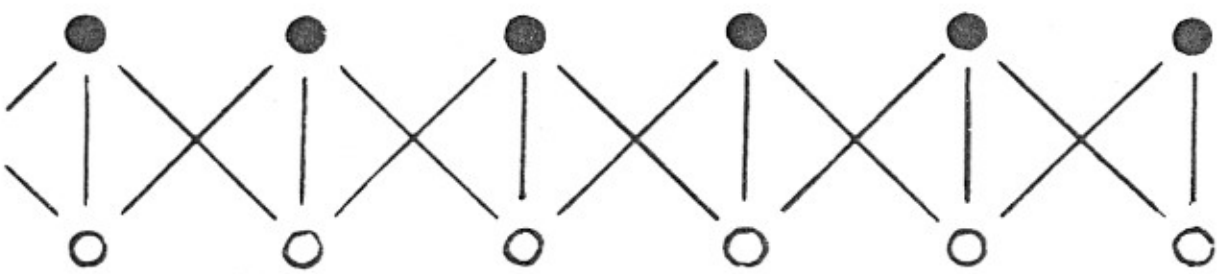

Fig. 1.2 Dependences between automata in Example 1.4a.

Table 1.2

r	4	3	2	1	0
θ	$\dfrac{1}{1+\varepsilon^2}$	$\dfrac{1}{1+\varepsilon}$	$\dfrac{1}{2}$	$\dfrac{\varepsilon}{1+\varepsilon}$	$\dfrac{\varepsilon^2}{1+\varepsilon^2}$

small enough has at least two different invariant measures: one near the state 'all zeros' and the other near the state 'all ones'. In this sense we have 'presented' or 'found' two invariant measures of our system.

Examples 1.4a and 1.4b have the following property of 'reversibility': the transition operator from t to $t + 1$ and the back transition operator from $t + 1$ to t are one and the same, up to a shift of the lattice. We shall describe general conditions under which a system of n state automata is reversible. All these systems have a Gibbs invariant measure, and the corresponding joint distribution at times t and $t + 1$ is also Gibbs and has some pair potential $\{\varphi_{ij}(x_i^t, x_j^{t+1})\}$, where (i,j) are all possible pairs of neighbours. This joint distribution is convenient to start with if you want to construct and investigate systems of this sort.

The next example is analogous to Example 1.2; it illustrates another way to prove non-ergodicity of non-degenerate systems [86,87].

Example 1.5 Automata are in integer points (i,j) of a plane. At every unit of time two following transformations occur.

First, the state of every automaton, having been $x_{(i,j)}^t$ at time t, becomes a certain function of its neighbours' states:

$$f(x_{(i,j)+u_1}^t, \ldots, x_{(i,j)+u_R}^t).$$

Let us assume that every automaton has only two states 0 and 1 and the function f takes only two values 0 and 1. Moreover, assume the function f to be monotone, that is, $a_1 \leqslant b_1, \ldots, a_R \leqslant b_R \Rightarrow f(a_1, \ldots, a_R) \leqslant f(b_1, \ldots, b_R)$.

The second transformation consists in that every automaton 'makes an error', that is, changes its state with probability ε independently from other automata and from all past events.

It remains to consider some natural R, some integer vectors $u_1 \ldots, u_R$ and some monotone function f: $\{0;1\}^R \to \{0;1\}$ to define a concrete system. We shall do it, but first we shall say something about our method of proving non-ergodicity of the system with small ε and about the conditions under which this method can work.

Non-ergodicity will result from two following facts:

(a) starting from the state 'all zeros', the system has a small proportion of ones at all times.
(b) starting from the state 'all ones', the system has a small proportion of zeros at all times.

Let us discuss how to prove (a) – because (b) can be proved analogously. It is clear that the proportion of ones will only increase if we change the second transformation in the following way: only errors from 0 to 1 occur, that is, automata having got into the state 0 make errors with

probability ε, but automata having got into the state 1 make no errors. Thus we have obtained a system which is similar to the Example 1.2 system in that their second transformations are alike. So we want to make their first transformations alike too. The first transformation in Example 1.2 diminished arrays of ones, and it was essential to resist the random generation of ones by the second transformation. So the first transformation in the present case has to do the same. Now we are ready to make things concrete.

Neighbours of (i,j) are

$$(i,j),\ (i-1,j),\ (i,j-1),\ (i-1,j-1)$$

and the value of f at (i,j) is

$$x_{(i,j-1)}x_{(i,j)} \vee x_{(i-1,j-1)}x_{(i-1,j)},$$

where $\vee$ means logical disjunction:

$$0\vee 0 = 0;\ 0\vee 1 = 1\vee 0 = 1\vee 1 = 1.$$

Let us explain why this system does work in the desired way. Let 'island' mean any state of our system having a finite (or empty) set of ones. It is essential that for any island there is such a t that t-multiple application of our first transformation turns this island into the state 'all zeros'. Moreover, this transformation erodes islands in a specific way: they are flattened along the j-axis, that is, the range of the values of j in the set of points occupied by ones is certain to diminish at every unit of time. This is the main point in the proof of the fact (a). The fact (b) is based on the analogous property of our transformation with respect to 'island of zeros in the sea of ones', that is, state having a finite number of zeros.

Note that the first transformation in our Example 1.1b (voting with three neighbours in a plane) erodes all islands of ones and all islands of zeros too. However, it does it in a different way, diminishing triangles whose sides are directed along i and j axes and lines $i + j = \text{const}$. This demands a more complicated proof of non-ergodicity, which, however, exists.

As to the Example 1.1c system, its non-ergodicity (if it takes place) seems to demand a quite new method of proof because its first (deterministic) transformation does not erode islands at all: neither of ones, nor of zeros.

Analogous problems may be considered for continuous time systems too [48], or for continuous space systems, or for a continuous set of a single-automaton states systems. In some way such systems are like discrete ones, but there are differences which present specific difficulties. We shall briefly mention these systems when treating analogous discrete systems in our survey.

Chapter 2
Definitions and notations

This chapter is unlike the previous one. It contains definitions and primary propositions general enough to serve most developments in this survey. First of all, let us define the system state space in a way suitable for all dimensions and all finite sets of states of a single automaton.

The state space. Let V be a countable or finite set. Every element (automaton) $h \in V$ has a finite set X_h termed its 'state space'. The system state space is their product:

$$X = \prod_{h \in V} X_h.$$

States of the system are elements $x \in X$ of the form

$$x = (x_h),\ h \in V.$$

We term x_h the hth coordinate of x.

For any non-empty $K \subset V$ we denote

$$X_K = \prod_{h \in K} X_h,$$

any element of which $x_K \in X_K$ may be viewed as the restriction of $x \in X$ onto K, that is, the image of x in the canonical projection $X \to X_K$.

As a rule, we assume that there is some transitive group G of permutations $g\colon V \to V$, in which case V is termed homogeneous (or G-homogeneous). This is convenient if all X_h are identical: $X_h = X_0$ for all $h \in V$. This assumed, any permutation $g\colon V \to V$ defines a map (which we denote by the same letter) $g\colon X \to X$ according to the rule:

$$(xg)_h = x_{g(h)} \quad \text{for all} \quad h \in V,\ x \in X.$$

(We write transformations on the right-hand side: $x \to xg$. The convenience of this will become clear.)

Elements of X_0 will be denoted by numbers in most cases:

$$X_0 = \{0;1\} \quad \text{or} \quad X_0 = \{1,2,\ldots,n\}.$$

For any $s \in X_0$ the state 'all x_h equal s' will be denoted $\mathbf{s}$.

For example, take $V = G = \mathbb{Z}$, where any $g \in G$ acts as a shift $h \to h + g$. Take also $X_0 = \{0;1\}$. In this case the space $X = \{0;1\}^{\mathbb{Z}}$ is the set of all both-sides-infinite sequences of zeros and ones. Any $g \in \mathbb{Z}$ shifts any sequence in question g units to the right (to the left if $g < 0$).

We equip our product $X = \Pi X_h$ with the discrete-factors product topology in which

$$\lim_{n\to\infty} x^{(n)} = x$$

means that for every $h \in V$ there is such N that

$$x_h^{(n)} = x_h \quad \text{if} \quad n \geqslant N.$$

Since all X_h are finite, X is compact in this topology.

Let us have two state spaces

$$x = \prod_{h\in V} X_h \quad \text{and} \quad X' = \prod_{i\in W} X'_i$$

A mapping $D\colon X \to X'$ is continuous in our topology if and only if every component x'_i of x' depends only on a finite set of components of $x \in X$. We term any continuous $D\colon X \to X'$ a deterministic transition operator from X to X'.

This survey is about stochastic transition operators, of which deterministic ones are a special case. Stochastic transition operators can be seen as linear operators acting upon either measures or real-valued functions on X. Both should be introduced now.

Measures. Cylinder sets are subsets of X having the form

$$\{x\colon x_K \in B\}$$

where B is any subset of X_K, K being any finite subset of V. A measure on X means any probabilistic measure μ (that is, $\mu(X) = 1$) on the σ-algebra generated by all cylinder sets. According to the well-known Kolmogorov theorem, any additive function on the algebra of cylinder sets uniquely extends to X. Moreover, we need not consider all cylinder sets; to define μ it is sufficient to assign its values consistently on elementary cylinder sets of the form

$$\{x\colon x_K = y_K\}$$

where $y_K \in X_K$. We shall denote briefly

$$\mu(\{x\colon x_K = y_K\}) = \mu(y_K).$$

Let us term a family of cylinder sets 'full' if a measure is uniquely defined by its values on the sets of this family. We have just said that the family of elementary cylinder sets is full. These are three other examples of full families.

(a) Let $X = \{0;1\}^V$. In this case a full family consists of sets of the following form:

$$\{x\colon x_{h_1}=1,\ldots,x_{h_m}=1\},$$

where $\{h_1,\ldots,h_m\}$ is any finite subset of V. The value of μ at this set is denoted briefly

$$\mu(x_{h_1} = \ldots x_{h_m} = 1).$$

(b) Let $X = \{0;1\}^{\mathbb{Z}}$. In this case a full family consists of sets of the form

$$\{x\colon x_i=0,x_{i+1}=s_1,\ldots,x_{i+m-1}=s_{m-1},x_{i+m}=0\},$$

where $s_1, \ldots, s_{m-1} \in \{0;1\}$ and $i, m \in \mathbb{Z}$, $m \geqslant 0$.

(c) Let $X = X_0^{\mathbb{Z}}$. In this case a full family consists of sets of the form

$$\{x\colon x_i=s_0,x_{i+1}=s_1,\ldots,x_{i+m}=s_m\},$$

where $s_0, \ldots, s_m \in X_0$ and $i, m \in \mathbb{Z}$, $m \geqslant 0$.

Let $\mathcal{M}(X)$ stand for the set of all normed measures on X. For any μ, $\nu \in \mathcal{M}(X)$ their mixture also belongs to $\mathcal{M}(X)$:

$$\alpha\mu + (1 - \alpha)\nu \in \mathcal{M}(X) \text{ for any } \alpha,\ 0 \leqslant \alpha \leqslant 1.$$

Hence $\mathcal{M}(X)$ is convex.

Now let us introduce a weak topology in $\mathcal{M}(X)$ by assuming a sequence of measures convergent if it converges in every cylinder set. This makes $\mathcal{M}(X)$ compact. The symbol δ_x stands for the measure concentrated in one point $x \in X$. The mapping $X \to \mathcal{M}(X)$ defined by $x \to \delta_x$ is continuous.

We call a measure μ homogeneous if $\mu(Ag) = \mu(A)$ for all cylinder A and all $g \in G$. The set $\mathcal{M}_G(X)$ of homogeneous measures is convex and closed in $\mathcal{M}(X)$.

We assume the image of a measure in a continuous or deterministic mapping and the canonical projection of a measure in the mapping $X \to X_K$ to be defined in the natural way.

We term a measure on a product space $X = \Pi X_h$ Bernoulli or independent if it equals the product of its projections onto single X_h sets.

Operators and their graphs. Let us have two spaces

$$X = \prod_{h \in V} X_h \quad \text{and} \quad X' = \prod_{i \in W} X'_i.$$

In particular, X and X' may be identical and correspond to times t and $t + 1$ of a system's work.

The following definition gives a general framework for our examples in Chapter 1, where every automaton's state depended in a random way on a finite set of its neighbours at the previous time.

Definition 2.1 A local operator P from X to X' stands for a linear operator P: $\mathscr{M}(X) \to \mathscr{M}(X')$ such that for any finite $K \subset W$ the projection of μP to X'_K depends only on the projection of μ to some $X_{U(K)}$ where $U(K)$ is a finite subset of V.

Let us write xP for $\delta_x P$ and say that P transforms x into it. A local operator is defined by its action on all $x \in X$. Definition 2.1 provides continuity of all mappings $\mu \to \mu P$ and $x \to xP$. Local operators will be our central theme. So we shall speak just of operators P from X to X', implying the omitted word 'local' everywhere.

Definition 2.2 An operator P is termed independent if the measure xP is independent (that is Bernoulli) for every x.

Definition 2.3 Suppose V and V' are G-homogeneous. An operator P from X_0^V to $Y_0^{V'}$ is termed G-homogeneous if $g \circ P = P \circ g$ for all $g \in G$.

Definition 2.4 An operator P from X to X' is termed non-degenerate if $\mu P(A) > 0$ for any $\mu \in \mathscr{M}(X)$ and any non-empty cylinder $A \subset X'$.

The bulk of our survey is about independent homogeneous operators from X to X, and we denote their totality by $\mathscr{P}(X)$. The independence condition is essential for many proofs. The homogeneity condition is not so essential but without it all our constructions would be more cumbersome. Let $\mathscr{P}_G(X)$ stand for the set of G-homogeneous elements of $\mathscr{P}(X)$.

In all our cases the neighbourhood relations can and will be constructed in such a way that

$$U(K) = \bigcup_{i \in K} U(i)$$

for all K. This allows us to explain any operator functioning with a suitable oriented graph $\Gamma = \Gamma(V, \mathscr{U})$, $\mathscr{U} = \{U(h), h \in V\}$. Its set of vertices is V and to any $h \in V$ arrows (directed bonds) come from all elements of $U(h)$. Say, Examples 1.2 and 1.3 where $V = \mathbb{Z}$, $U(h) = \{h-1; h\}$ have the graph shown in Figure 2.1. Generally, for a homogeneous operator, its graph is homogeneous too, that is,

$$U(g(h)) = g(U(h))$$

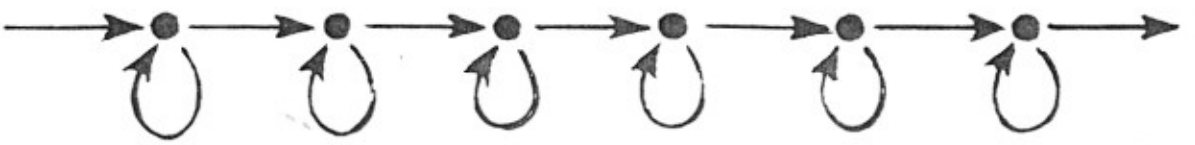

Fig. 2.1 Graph Γ_1.

We denote also:

$$U^m(h) = \underbrace{U(U(\ldots(U(h))\ldots))}_{m}$$

and

$$U^\infty(h) = \bigcup_{m=0}^{\infty} U^m(h).$$

The set $U^m(h)$ consists of those vertices of Γ from which there is an oriented path to h consisting of m bonds.

Mostly our operators' graphs will have the following property. We call Γ strongly connected if for any two of its vertices i and j there is such a vertex k that

$$\{i,j\} \subset U^\infty(k).$$

The following property makes graphs especially easy for investigation. We call Γ 'tree-like' if two different paths coming from a vertex never meet before; this amounts to $\{i,j\} \subset U(h)$, $i \neq j \Rightarrow U^\infty(i) \cap U^\infty(j) = \varnothing$. Evidently, an independent operator having a tree-like Γ transforms any Bernoulli measure into a Bernoulli measure.

Functions. Denote by $F = F(X)$ the totality of real-valued functions on X, of which every one actually depends only on a finite set of coordinates x_h: $\varphi(x) = \varphi(x_K)$ where K is finite. Often we shall identify a function $\varphi \in F(X)$ and its restriction to X_K on which it actually depends, as has been done in the last equation. In this sense we may write

$$F(X) = \bigcup_{K \subset V} F(X_K)$$

where all K are finite.

In particular, $F(X_h)$ is a linear space, whose dimension is $|X_h|$. One of its bases consists of the following functions:

$$\chi_h^s(x) = \begin{cases} 1 & \text{if } x_h = s, \\ 0 & \text{otherwise,} \end{cases}$$

where $h \in V$ is fixed, s runs through X_h.

These functions are convenient because the whole space $F(X)$ consists of finite linear combinations of finite products of them. Let $F^+(X) \subset$

$F(X)$ stand for non-negative-valued functions from $F(X)$. They are the same linear combinations with non-negative coefficients.

We equip $F(X)$ with a norm:

$$\|\varphi\| = \sup_{x \in X} |\varphi(x)|.$$

For any linear operator P: $M(X) \to M(X')$ there is the conjugate operator P: $F(X') \to F(X)$. As you see, we denote them by one and the same letter. This allows us to reword our Definitions 2.1–4 in the following way:

Definition 2.1′ A (local transitional) operator P from X to X' stands for a linear operator

$$P\colon F(X') \to F(X)$$

satisfying conditions

$$P(F^+(X')) \subset F^+(X) \quad \text{and} \quad P(1_{X'}) = 1_X,$$

where $1_{X'}$ and 1_X are function on X' and X identically equal to 1.

Definition 2.2′ An operator P from X to X' is called independent if

$$P(\varphi\psi) = P\varphi \cdot P\psi$$

for any such

$$\varphi \in F(X'_K) \quad \text{and} \quad \psi \in F(X'_L)$$

that $K \cap L = \varnothing$ (where K and L are finite subsets of V').

This multiplicative interpretation makes it useful to treat an independent operator as acting on functions, instead of measures.

For any continuous mapping f: $X \to X$ let T_f stand for the linear operator defined by

$$T_f(\varphi) = \varphi \circ f$$

where $\varphi \in F(X)$. In particular, this defines Tg for all $g \in G$.

Definition 2.3′ Let V and V' be G-homogeneous. An operator P from X_0^V to $Y_0^{V'}$ is termed G-homogeneous if

$$Tg \circ P = P \circ Tg$$

for all $g \in G$.

Definition 2.4′ An operator P from X to X' is termed non-degenerate if for any $\varphi \in F^+(X')$ the condition $\varphi \not\equiv 0$ implies $(P\varphi)(x) > 0$ for any $x \in X$.

Let us denote

$$\mu\varphi = \int_X \varphi(x)d\mu(x).$$

Then by definition of the conjugate operator,

$$(\mu P)\varphi = \mu(P\varphi)$$

which warrants writing this simply as $\mu P\varphi$.

Now we see the point of our decision to write operators on the left-hand side of functions and on the right-hand side of measures. In the same way we can write composition of operators: suppose P is an operator from X to X' and Q is an operator from X' to X''. Then PQ is an operator from X to X'' and for any $\mu \in \mathcal{M}(X)$ and $\varphi \in F(X'')$

$$(\mu PQ)\varphi = (\mu P)(Q\varphi) = \mu(PQ\varphi),$$

which may be written as just $\mu PQ\varphi$.

Often we shall define an independent operator P from X to $X = \prod_{h\in V} X_h$ in the following way. For every h we fix its neighbourhood $U(h) \subset V$ and transition probabilities

$$\theta_h(y_h|x_{U(h)}) \in [0;1]$$

for all $y_h \in X_h$, $x_{U(h)} \in X_{U(h)}$, which are subject to condition

$$\sum_{y_h} \theta_h(y_h|x_{U(h)}) = 1 \quad \text{for all} \quad x_{U(h)}.$$

Now our operator is fixed. For any measure $\mu \in \mathcal{M}(X)$, any $K \subset V$ and any $y_K \in X_K$,

$$\mu P(y_K) = \sum_{x_{U(K)}} \mu(x_{U(K)}) \prod_{h\in K} \theta_h(y_h|x_{U(h)}). \tag{2.1}$$

For any $\varphi \in F(X_K)$ image $P\varphi \in F(X_{U(K)})$ is defined by

$$P\varphi(x) = \sum_{y_K} \varphi(y_K) \prod_{h\in K} \theta_h(y_h|x_{U(h)}). \tag{2.2}$$

If V is G-homogeneous, G-homogeneity of P amounts to the following consistency of transition probabilities:

$$\theta_{g(h)}(y_{g(h)}|x_{U(g(h))}) = \theta_h((yg)_h|(xg)_{U(h)}).$$

In this case, defining all the transition probabilities boils down to defining them at one place, say $0 \in V$:

$$\theta_0(y_0|x_{U(0)}) \text{ or } \theta^{y_0}_{x_{U(0)}}.$$

Sometimes we need to define an operator on $X = \Pi X_h$ in the following

way. First, we have an auxiliary space $\Omega = \prod_{h \in V} \Omega_h$ where all Ω_h are finite, and a measure ν on Ω. Second, we have a mapping

$$f\colon \Omega \times X \to X$$

defined by certain mappings for all $h \in V$:

$$f_h\colon \Omega_h \times X_{U(h)} \to X_h$$

where $U(h)$ are given for all $h \in V$.

Now we can define P by saying that μP is induced by the measure $\nu \times \mu$ on the space $\Omega \times X$ with the mapping f. If ν is independent, that is, $\nu = \prod_{h \in V} \nu_h$ where $\nu_h \in \mathcal{M}(\Omega_h)$, the operator P turns out independent and its transition probabilities are

$$\theta_h(y_h | x_{U(h)}) = \Sigma \nu_h(\omega_h)$$

where the summing is over all ω_h such that

$$f_h(\omega_h, x_{U(h)}) = y_h.$$

The trivial kind of independent operators are those for which $U(h)$ consists only of h for every h. This means that automata do not interact at all, every one behaving as a Markov chain of its own. Let us term such an operator a 'noise'. We shall often use the following two special noises acting on $\{0;1\}^V$ where ε is the parameter, $0 \leqslant \varepsilon \leqslant 1$:

(a) the noise S_ε defined by $\theta_0^1 = \varepsilon$, $\theta_1^1 = 1$, which means that every automaton goes from 0 to 1 with probability ε and remains in 1 with probability 1;
(b) the noise S_ε^+ defined by $\theta_0^1 = \theta_1^0 = \varepsilon$, which means that every automaton always changes its state with probability ε.

Note that composition of an independent operator and a noise is another independent operator with the same neighbours. In particular, every operator in our Examples 1.1, 1.2 and 1.3 was actually described as a composition of a deterministic operator and a noise.

Our construction with its auxiliary space Ω is very suitable for representing these operators. For any deterministic operator D the composition DS_ε is representable with $\Omega_h = \{0;1\}$, $\nu_h(1) = \varepsilon$ and $f_h(\omega_h, x) = D_h(x) \vee \omega_h$, and the composition DS_ε^+ is representable analogously, only

$$f_h(\omega_h, x) = D_h(x) + \omega_h \pmod 2.$$

(The cases $\omega_h = 1$ correspond to errors of automata.)

It is possible to represent all independent operators, squeezing all the randomness into an independent measure on an auxiliary space Ω, but in the general case it seems to be more cumbersome than useful.

Invariant measures and ergodicity. Let us call a measure invariant for an operator P if $\mu = \mu P$.

Proposition 2.5 Let an operator P map some convex compact $C \subset \mathscr{M}(X)$ into itself: $CP \subset C$. Then P has at least one invariant measure $\mu \in C$.

This is a version of the well-known fixed-point theorem. The reader can hit on its proof having taken any initial $\mu_0 \in C$ and proving $\mu = \mu P$ for $\mu = \lim_{n\to\infty} \frac{1}{n}(\mu_0 + \mu_0 P + \ldots + \mu_0 P^{n-1})$. Now we can take $C = \mathscr{M}(X)$ and obtain:

Corollary 2.6 Any operator has an invariant measure. Any homogeneous operator has a homogeneous invariant measure. Hence the set $\bigcap_{t=0}^{\infty} \mathscr{M}P^t$ is non-empty as it contains all the invariant measures.

Definition 2.7 Term an operator $P \in \mathscr{P}(X)$ ergodic if the set $\bigcap_{t=0}^{\infty} \mathscr{M}P^t$ consists of one point.

The following is its equivalent:

Definition 2.7′ Term an operator $P \in \mathscr{P}(X)$ ergodic if for any $\varphi \in F(X)$ the sequence $P^t\varphi$ uniformly converges to a constant with $t \to \infty$.

To prove equivalence of Definitions 2.7 and 2.7′ note that both are equivalent to the following: For any $\varphi \in F$ the contracting sequence of intervals

$$\Delta_t(\varphi) = \{\mu P^t \varphi : \mu \in \mathscr{M}\} \subset \mathbb{R}, \quad t = 1, 2, 3, \ldots$$

has just one common point.

Clearly, ergodicity implies uniqueness of invariant measure, but generally not vice versa. (The simplest example: V is finite and consists of one point h; so we have just one automaton, which has two states 0 and 1 and passes from each state to the other with probability 1. This system is non-ergodic, but the only invariant measure is $(\frac{1}{2},\frac{1}{2})$.) In fact, however, all the proofs of non-ergodicity in our survey are proofs of non-uniqueness of invariant measure. Often these proofs use another corollary of Proposition 2.5:

Corollary 2.8 P is an operator on $\mathscr{M}(X)$, C_1 is a convex compact in $\mathscr{M}(X)$ and there is such a measure μ that $\mu P^t \in C_1$ for all natural t, then P has an invariant measure in C_1.

To prove Corollary 2.8, one has only to apply Proposition 2.5 where C should stand for the closing of the convex hull of the union of μP^t for all natural t.

Let us consider a simple example of a non-ergodic non-degenerate operator. Let $\Gamma = \Gamma(V, \mathcal{U})$ be a homogeneous tree-like graph with $|U(h)| = 3$. Let every automaton have two states 0 and 1 and choose $P \in \mathcal{P}(\{0,1\}^V)$ in the form $P_\varepsilon = DS_\varepsilon^+$ where D is a deterministic 'majority voting' operator:

$$(Dx)_h = 1, \quad \text{if} \quad \sum_{k \in U(h)} x_k \geqslant 2,$$

$$(Dx)_h = 0, \quad \text{if} \quad \sum_{k \in U(h)} x_k \leqslant 1.$$

Let ν_γ stand for the homogeneous Bernoulli measure on $\{0;1\}^V$ where γ is the proportion of ones:

$$\nu_\gamma(x_h = 1) = \gamma \quad \text{for all} \quad h \in V.$$

Any homogeneous operator on a tree-like graph transforms a homogeneous independent measure into another of the same sort. In the present case

$$\nu_\gamma P_\varepsilon = \nu_{f(\gamma,\varepsilon)}$$

where

$$f(\gamma,\varepsilon) = (1 - 2\varepsilon)(\gamma^3 + 3(1 - \gamma)\gamma^2) + \varepsilon.$$

Now we have only to solve the equation

$$f(\gamma,\varepsilon) = \gamma$$

to obtain all homogeneous independent invariant measures. Solutions of this equation are $\gamma = \frac{1}{2}$ and $\gamma = \frac{1}{2}\{1 \pm \sqrt{[(1 - 6\varepsilon)/(1 - 2\varepsilon)]}\}$, the latter being valid for $\varepsilon < \frac{1}{6}$. Thus, with $\varepsilon < \frac{1}{6}$ we have three different invariant measures of one operator P.

Note. The same formulae allow another interpretation. We may assume that $P_\varepsilon = DS_\varepsilon^+$ where Γ is any homogeneous graph, but all automata are randomly mixed after every application of P_ε.

This mixing transforms any homogeneous measure into an independent measure having the same proportion of ones. Thus, all homogeneous invariant measures of a system with mixing are independent and their investigation boils down to examination of iterations of one function f: $[0;1] \to [0;1]$ where

$$f(\gamma) = (\nu_\gamma)P(x_h = 1).$$

The behaviour of operators with intermixing may be seen as an approx-

imation of the chaotic behaviour of operators, about which there is an extensive literature, including [67].

Evolution graphs. Remember that we are interested in the temporal behaviour of automata systems. As before, x_h^t stands for the state of the hth automaton at time t. Let $\Gamma(V, \mathscr{U})$ stand for the graph of our operator P. We know that any x_h^{t+1} depends in a probabilistic way on $x_{U(h)}^t$, which depends on $x_{U^2(h)}^{t-1}$, which depends on $x_{U^3(h)}^{t-2}$, and so on. To represent these dependences in a more suitable way we introduce

$$\tilde{X} = X^{\mathbb{Z}_+} = \{(x_h^t), h \in V, t \in \mathbb{Z}_+\}.$$

Thus, an element of $\tilde{X}$ is given by states of the system at all the times 0, 1, 2, ... Now we introduce the oriented graph $\tilde{\Gamma}$ whose set of vertices is

$$\tilde{V} = \{(h,t), h \in V, t \in \mathbb{Z}_+\}.$$

To every vertex $(h,t+1)$ of $\tilde{\Gamma}$ arrows (oriented bonds) come from all (k,t), where $k \in U(h)$, and the corresponding neighbourhood is denoted

$$\tilde{U}((h,t+1)) = \{(k,t), k \in U(h)\}.$$

We term $\tilde{\Gamma}$ the evolution graph of Γ. The same term is also used if t runs over a finite range. We term a set $K \subset \tilde{V}$ pyramidal if

$$v \in K \Rightarrow \tilde{U}(v) \subset K.$$

Of course, any operator P and initial measure μ (given at time $t = 0$) define a certain measure $\tilde{\mu}$ ou $\tilde{X}$, that is, the joint probabilistic distribution for all the times 0, 1, 2, ... We shall call $\tilde{\mu}$ evolution measure. For a local independent operator P having $\theta(y_h | x_{U(h)})$ parameters the value of evolution measure $\tilde{\mu}$ on a cylinder set X_K for K pyramidal is equal to

$$\tilde{\mu}(x_K) = \mu(x_{K \cap V^0}) \prod_{v \in K \setminus V^0} \theta_{h(v)}(x_v | x_{\tilde{U}(v)}),$$

where $V^0 = \{(h,0), h \in V\} \subset \tilde{V}$ and $h(v)$ is the first component of $v = (h,t) \in \tilde{V}$. These values define $\tilde{\mu}$, that is, the totality of pyramidal sets is full.

Monotonicity. Sometimes it is convenient to introduce a partial order in X and $\mathscr{M}(X)$. First every X_h has to be ordered. If elements of X_h are denoted by numbers, it is assumed that

$$p \prec q \Leftrightarrow p \leqslant q.$$

In particular, in the case $X_h = \{0;1\}$ we always assume $0 \prec 1$.

In any event, suppose every X_h ordered. Then we partially order X assuming $(x_h) \prec (y_h)$ if $x_h \prec y_h$ for all $h \in V$. We term a real-valued function φ on X monotone if

$$x \prec y \Rightarrow \varphi(x) \leqslant \varphi(y).$$

We order $\mathscr{M}(X)$ too by assuming that $\mu \prec \nu$ if $\mu\varphi \leqslant \nu\varphi$ for any monotone $\varphi \in F(X)$.

We term a set $C \subset X$ full-above if

$$x \in C, x \prec y \Rightarrow y \in C.$$

It is easy to see that $\mu \prec \nu$ if and only if $\mu(C) \leqslant \nu(C)$ for all full-above C.

Proposition 2.9 Let $X = \prod_{h \in V} X_h$. Let μ and ν be independent measures on X, that is,

$$\mu = \prod_{h \in V} \mu_h, \ \nu = \prod_{h \in V} \nu_h.$$

Then $\mu \prec \nu$ if and only if $\mu_h \prec \nu_h$ for all $h \in V$.

One half of this assertion is evident. Let us prove the other: suppose $\mu_h \prec \nu_h$ for all h and prove $\mu \prec \nu$, that is, prove $\mu\varphi \leqslant \nu\varphi$ for any $\varphi \in F(X)$. First assume that $\varphi \in F(X_K)$ where K is a finite subset of V. In this case it is sufficient to prove that $\mu_K \prec \nu_K$ where

$$\mu_K = \prod_{h \in K} \mu_h, \ \nu_K = \prod_{h \in K} \nu_h.$$

because $\mu\varphi = \mu_K\varphi$ and $\nu\varphi = \nu_K\varphi$. This can be proved by induction over the number of elements in K. Essentially the induction step is analogous to the proof for K having two elements. Finally, we prove $\mu\varphi \leqslant \nu\varphi$ for any φ approximating it with functions of a finite set of arguments.

Definition 2.10 [52]. Let X and X' be partially ordered. We term an operator P from X to X' monotone if

$$x \prec y \Rightarrow xP \prec yP.$$

Equivalently, we may term P monotone if

$$\mu \prec \nu \Rightarrow \mu P \prec \nu P.$$

Equivalently, we may term P monotone if for any monotone φ the function $P\varphi$ is monotone too.

Definition 2.11 Let P and Q be two operators from X to X' and X' be partially ordered. Let $P \prec Q$ mean that $xP \prec xQ$ for any $x \in X$. Equivalently, $P \prec Q$ means that $\mu P \prec \mu Q$ for any $\mu \in \mathscr{M}(X)$.

Proposition 2.12 An operator P on $\{0;1\}^V$ having parameters $\theta_h(y_h | x_{U(h)})$ is monotone if and only if

$$x_{U(h)} \prec x'_{U(h)} \Rightarrow \theta_h(1 | x_{U(h)}) \leqslant \theta_h(1 | x'_{U(h)}).$$

Proposition 2.13 Let P and P' be operators on $\{0;1\}^V$ having parameters $\theta_h(y_h|x_{U(h)})$ and $\theta'_h(y_h|x_{U(h)})$ respectively. Then $P \prec P'$ if and only if for all $x_{U(h)}$

$$\theta_h(1|x_{U(h)}) \leqslant \theta'_h(1|x_{U(h)}).$$

Proposition 2.14 Let us have operators P from X to X, Q from X' to X' and R from X to X' and let X' be partially ordered, Q be monotone, and $RQ \prec PR$. Then for all natural t

$$RQ^t \prec P^t R$$

Putting here $X = X'$ and R the identity operator, we obtain:

Corollary 2.15 Let P and Q be operators on X, let $P \prec Q$ and let at least one of P and Q be monotone. Then $P^t \prec Q^t$ for all natural t.

The set $\mathscr{P}_1$. Let $\mathscr{P}_1 = \mathscr{P}_1(\{0,1\}^V)$ stand for the set of those independent homogeneous operators P on $X = \{0;1\}^V$ which conserve the measure δ_1 concentrated in the point 'all ones': $\mathbf{1}P = \delta_1$. Having δ_1 as an invariant measure, these operators are ergodic if they attract all X to this measure in the limit $t \to \infty$. To consider this problem more closely, we denote

$$\rho_K^t(Q) = \mathbf{0}Q^t(\mathbf{1}_K)$$

where $K \subset V$, t is time, $\mathbf{0}$ is the state 'all zeros', $\mathbf{1}_K$ is the cylinder set $\{x\colon x_h = 1 \text{ for all } h \in K\}$. If Q is monotone (in which case we shall use this notation), the sequence $\rho_K^t(Q)$ is non-decreasing in t for any K, whence all the limits exist:

$$\rho_K(Q) = \lim_{t\to\infty} \rho_K^t(Q).$$

So the limit measure exists:

$$\mu_Q = \lim_{t\to\infty} \mathbf{0}\, Q^t$$

and is invariant for Q.

Proposition 2.16 For any monotone $Q \in \mathscr{P}_1$ just one of the two following cases takes place:

(a) Either Q is ergodic and $\mu_Q = \delta_1$
(b) Or Q is non-ergodic and $\mu_Q \neq \delta_1$ and

$$\rho_K(Q) \leqslant \mu_Q(1_h) < 1$$

for all $K \neq \varnothing$, $h \in V$.

It is easy to prove that some operators of $\mathscr{P}_1$ are ergodic: namely those which have large enough parameters θ_z^1, $z \in U(h)$.

Proposition 2.17 Let an operator $P \in \mathscr{P}_1(\{0,1\}^V)$ have parameters $\theta^y_{x_1 \ldots x_R}$ where $R = |U(h)|$. Suppose that there exist such non-negative $\varkappa_1, \ldots, \varkappa_R$ that $\sum_{i=1}^{R} \varkappa_i < 1$ and for any $x_1, \ldots x_R \in \{0;1\}$

$$\theta^0_{x_1 \ldots x_R} \leqslant \sum_{i=1}^{R} \varkappa_i (1 - x_i).$$

Then P is ergodic.

Proof. Let μ be any initial measure and $\tilde{\mu}$ be the related evolution measure. Then

$$\mu P^{t+1}(x_h = 0) = \tilde{\mu}(x_h^{t+1} = 0) = \sum \theta^0_{x^t_{U(h)}} \tilde{\mu}(x^t_{U(h)}) \leqslant$$

$$\leqslant \sum \tilde{\mu}(x^t_{U(h)}) \sum_{i=1}^{R} \varkappa_i (1 - x_i^t) \leqslant$$

$$\leqslant \sum_{i=1}^{R} \varkappa_i \sum_{x^t_{U(h)} : x_i^t = 0} \tilde{\mu}(x^t_{U(h)}) = \sum_{i=1}^{R} \varkappa_i \tilde{\mu}(x_i^t = 0).$$

Let us denote $\sigma_t = \sup_{h \in V} \tilde{\mu}(x_h^t = 0)$,

$$\varkappa = \varkappa_1 + \ldots + \varkappa_R < 1.$$

Then $\sigma_{t+1} \leqslant \varkappa \sigma_t$ whence $\sigma_t \leqslant \varkappa^t \to 0$ q.e.d.

Consider a simple example: an operator $P \in \mathscr{P}_1(\{0,1\}^V)$ on the simplest graph Γ_1:

$$V = \mathbb{Z};\ U(h) = \{h-1, h\},\ h \in \mathbb{Z},$$

and denote

$$\theta(x_h^{t+1} = 1 \,|\, x_{h-1}^t, x_h^t) = \theta_{x_{h-1}^t x_h^t}.$$

In our case $\theta_{11} = 1$. Proposition 2.17 proves ergodicity in the case

$$\theta_{10} + \theta_{01} > 1 \quad \text{and} \quad \theta_{00} > 0.$$

In Chapter 5 we shall use another method to prove ergodicity for P monotone,

$$\theta_{00} + \theta_{01} + \theta_{10} > 1 \quad \text{and} \quad \theta_{00} > 0.$$

Consider another example: an operator $P \in \mathscr{P}_1$ on any graph. Of course, $\theta_{11 \ldots 1} = 1$. Let all others be: $\theta_z \equiv \varepsilon$ for all $z \neq \mathbf{1}$ (this is a 'percolation' operator). For this case our Proposition 2.17 proves ergodicity for $\varepsilon > 1 - \frac{1}{R}$, where $R = |U(h)|$.

For tree-like graphs this estimation is the best one. In fact, let us have an independent measure $\nu_\gamma \in \mathcal{M}(\{0;1\}^V)$ and a percolation operator P_ε on the tree-like graph $\Gamma(V, \mathcal{U})$ where $|U(h)| = R$. Then $\nu_\gamma P_\varepsilon = \nu_{f(\gamma,\varepsilon)}$ where $f(\gamma,\varepsilon) = (1 - \varepsilon)\gamma^R + \varepsilon$. It is easy to see that for $\varepsilon < 1 - \frac{1}{R}$ the equation $f(\gamma,\varepsilon) = \gamma$ has not only the trivial solution $\gamma = 1$ but another solution. (Because $f(0,\varepsilon) = \varepsilon > 0$ and $\frac{\partial f}{\partial \gamma}(1,\varepsilon) = R(1 - \varepsilon)$.) Thus, with $\varepsilon < 1 - \frac{1}{R}$ the operator P_ε is non-ergodic.

But for the graph Γ_1 the estimation $\varepsilon > 1 - \frac{1}{R}$ can be improved, and we shall do that in Chapter 8.