Generalized ϕ-Divergence and Frequency Analysis in Markov Chains

Dissertation

zur Erlangung des Doktorgrades
an der Naturwissenschaftlichen Fakultät
der Universität Salzburg
eingereicht von

Stefan Wegenkittl

4. Mai 1998

No one should be expected
to solve a math problem
that has a "twelve" in it.

– Peppermint Patty, Peanuts

It is difficult to predict,
especially the future.

– N. Bohr, old Danish proverb

To Doris.

It was a pleasure and a piece of good fortune for me to be able to write this thesis whilst being member of the pLab group at Salzburg.

Thank you, Peter, for introducing me to the world of pseudorandomness and for making all the collaborations possible. Thank you, Charly, Hannes, and Karin, for making Mathematics so adventurous and colourful, and – of course – for the good mood and music in our office.

The development of this work was accompanied by Prof. F. Österreicher from the very beginning. Thank you, Ferdinand, for the excellent teaching and your patience. The material and presentation owes much to your contributions!

Thank you, Max, for introducing me to the beauty of probability.

I gratefully acknowledge the advice, comments, corrections and encouragements of Prof. Pierre L'Ecuyer (who was the first to draw my attention to overlapping power divergences), Prof. Makoto Matsumoto (who invented the primary version of the gambling test), Dr. Grodzicki (who explained Bohr's model of the hydrogen atom to me), and my sister Susanne "Su" (who rendered the Salzburgian accent into something pretty much like British English).

This thesis would not have been started out without the generous support and understandning of my parents Renate and Willibald, and my brother Helmut and sister Susanne.
It would never have been finished without Doris always drawing my attention to the funny side of everything often taken so seriously. I dedicate this work to you.

Salzburg, Mai 4, 1998

This work was supported by the Austrian Science Foundation (FWF), project no. P11143-MAT led by Peter Hellekalek.

Contents

Chapter 1

Introduction

In 1900, K. Pearson [45] defined his famous goodness-of-fit statistic with respect to a multinomial model and derived the asymptotic chi-square distribution thereof. Various generalizations of this key result regarding statistical testing have been considered during the progress of the century. In 1949, Neyman [38] showed the asymptotics for the Neyman-modified statistic, in 1951, the log-likelihood statistic [26], and in 1959, the modified log-likelihood statistic [25] were introduced. These concepts for distance measures were unified in Csiszár's [8] approach in 1963, and independently in that of Ali and Silvey [1] in 1966: the concept of the φ-divergence allows to analyze various aspects common to all aforementioned test statistics at the same time including in particular the asymptotic distribution under multinomial distributed models.

In this thesis we extend the notion of φ-divergence from the multinomial model to the general case of asymptotically multivariate normal distributed data. We introduce a *generalized ϕ-divergence* $I_{\overline{\Sigma},\phi}$ and give conditions for the convergence in distribution to a chi-square distribution. As a particular interesting application of generalized ϕ-divergences we consider the long term behaviour of finite irreducible Markov Chains and thereby enlarge the class of so-called Serial tests.

Serial tests are goodness-of-fit statistics based on counting the occurrence of certain events and on comparing the resulting relative frequencies with the according expectations. In the case of independent events the vector of counters is multinomially distributed so that we can apply any of Csiszár's φ-divergence tests immediately.

In the case of m-dependent events, which arises, for example, when counting overlapping tuples of events, the matter tends to become more complicated. The covariance matrix Σ of the vector of counters differs from the multinomial case and the test statistic has to be modified accordingly in order to make its asymptotic distribution accessible. In 1938, Kendall and Smith discussed such Serial tests for the random sampling of numbers but they came to wrong conclusions for the overlapping case which was solved properly by Good [16] in 1953. The theory of quadratic forms in weak inverses $\overline{\Sigma}$, see e.g. Rao [47], gives a general framework for the necessary modifications of the Pearson statistic. It is based on the analysis of the covariance structure of the counters and applies to the general case of asymptotically multivariate normal data.

It seems quite natural to consider the notion of Markov chains in order to model Serial tests. In this context, the relative frequency count becomes the average occupancy time of any state of the chain. The chain allows to model a vast pool of applications in the field of linear control with noise.

Now, approaching the end of the 20'th century, we collect all these items and present a study where we connect the frequency analysis of Markov chains with our concept of generalized ϕ-divergences. Our motivation stems from a particular application of Serial tests in the field of pseudorandom number generation where we need the flexibility provided by the chain approach to imitate actual applications as far as possible. Usual Serial tests have not been able to reject huge-period linear generators even if such a generator had theoretically been proven to be highly defective. The example given in Section 5.2 indicates that generalized ϕ-divergence based Serial tests may overcome this limitation and should thus be viewed as promising candidate in the struggle for empirical evidence against defects in the workhorse generators of tomorrow's computer based stochastic simulation.

1.1 Outline and Summary

Chapter 2 discusses modelling with Markov chains. We introduce the notion of coupled-, overlapping- and higher-order chains and recall some known facts about the asymptotic behavior of cell-occupancy times. As a concrete example accompaning the theory we consider Bohr's model for the hydrogen atom. In Chapter 3 we construct Serial tests for the assessment of pseudorandom number generators by the aid of Markov chains and show

how Pearson's statistic and its generalizations may be applied to test certain properties of pseudorandom numbers.

Chapter 4 is devoted to the construction and discussion of the *generalized ϕ-divergence*. We introduce both, a general, and a special "easy-to-use" version of this measure and deduce the asymptotic distributions under multivariate normality. We will in particular consider the case of degenerate normal distributions. We also show, how Pearson's statistic, the φ-divergence, and quadratic forms may be obtained as special cases of the generalized ϕ-divergence. Chapter 5 contains examples and an application to the testing of pseudorandom number generators. Technical lemmata and proofs are deferred to the Appendix.

The reader who is mainly interested in the concept of generalized ϕ-divergence may skip Chapters 2 and 3. If the reader is interested in testing pseudorandom number generators on the other hand, we recommend the reading of Section 2.2, Chapter 3, and Section 5.2.

1.2 Notation and Concepts

We denote the set of natural, real, and complex numbers by $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{C}$, respectively, and define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and $\mathbb{N}_i := \{1, 2, \ldots, i\} \subset \mathbb{N}$ for $i \in \mathbb{N}$. For the finite index set $S = \mathbb{N}_m$, $m \in \mathbb{N}$, with cardinality $\#S = m$, we let $\mathbf{P} = (P_i)_{i \in S}$ denote the vector with the components P_i ordered by increasing indices. Vectors are notated in bold letters and are assumed to be row vectors in general, column vectors are denoted by adding a prime like in $\mathbf{P}'$. Let $\mathbb{N}_m^s$ denote the s-fold Cartesian product of $\mathbb{N}_m$ and assume lexicographical ordering such that $(1, \ldots, 1, 1) \prec (1, \ldots, 1, 2) \prec \ldots \prec (m, \ldots, m)$. This enables us to use $\mathbb{N}_m^s$ as index set for the components of a m^s-dimensional vector $\tilde{\mathbf{P}} = (\tilde{\mathbf{P}}_{\mathbf{i}})_{\mathbf{i} \in \mathbb{N}_m^s}$. Sequences with elements in a set S will be notated $(x_n)_{n \in \mathbb{N}}$, $x_n \in S$.
The transpose of a $k \times l$ matrix $A = (a_{ij})_{(i,j) \in \mathbb{N}_k \times \mathbb{N}_l}$ is denoted by A'. We further let $a_{ij}^{(n)}$, $(i, j) \in \mathbb{N}_m^2$, denote the elements of the n'th power A^n of the $m \times m$ square matrix A. Let $diag(\lambda_1, \ldots, \lambda_m)$ be the diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_m$ and let $I_m = diag(1, \ldots, 1)$ be the m-dimensional identity matrix. The symbol $\emptyset$ denotes the null vector $(0, \ldots, 0)$.

As to probabilistic contexts, we denote by $P[\mathcal{E}]$ the probability of the set or event $\mathcal{E}$. If X is a random variable, we denote its expectation and variance

by $E[X]$ and $V[X]$, respectively. If $\hat{\mathbf{P}} \in \mathbb{R}^m$ is a random vector, $E[\hat{\mathbf{P}}]$ denotes the vector of expectations and $V[\hat{\mathbf{P}}] = (v_{ij})_{(i,j)\in\mathbb{N}_m^2}$ denotes the covariance matrix with $v_{ii} = V[\hat{P}_i]$ and $v_{ij} = Cov[\hat{P}_i, \hat{P}_j]$ for $i \neq j$. If $(X_n)_{n\in\mathbb{N}}$ is a sequence of independently distributed random variables where each X_n obeys the same distribution, we call this an independent identically distributed (i.i.d.) sequence of random variables.

We use the symbol $\sim$ twofold. If X is a random variable, $X \sim Dist$ means that X obeys the distribution $Dist$ like in $X \sim \chi_k^2$. If $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are sequences of real numbers, $x_n \sim y_n$ means that x_n/y_n converges to 1 as n goes to infinity. Stochastic convergence is discussed in the appendix.

The reader is assumed to be familiar with the concept of random vectors in $\mathbb{R}^m$, with the Central Limit Theorem for sequences of such vectors, with the m-variate normal distribution $\mathcal{N}(\mu, \Sigma)$ with mean μ and covariance matrix Σ, and with the chi-square distribution χ_k^2 with k degrees of freedom.

Chapter 2

Modelling with Markov Chains

Consider the following example of a homogeneous stochastic process with discrete time and finite state space which is inspired by a simplified model of Bohr's hydrogen atom[1], see [21, Beispiel 1]. The physical model permits infinite many orbits for the electron and allows transitions among orbits at arbitrary time points. However, we identify the set of possible orbits of the electron with a *finite* set of states $S = \{1, \ldots, m\}$ for our purpose. Absorption or emission of photons cause the electron to jump from one state into another. We further assume that such transitions may take place only at certain instants $n \in \mathbb{N}_0$ of a discrete time unit. With a spectrograph, we are able to estimate the (hypothetical) probabilities p_{ij}, $(i,j) \in S^2$, for a transition from state i to state j. Homogeneity is the assumption that these probabilities do not depend on or vary with the time n. For the complete specification of the stochastic process $(X_n)_{n\in\mathbb{N}_0}$, where X_n is the state of the system at time point n, we need to provide a distribution of the initial state X_0, which we denote by $\mathbf{P}_0 = (P_{01}, \ldots, P_{0m})$. Here $P_{0i} = P[X_0 = i\,]$ denotes the probability of a start in state i, $i \in S$. There is no risk of confusion between the initial probabilities p_{0i} and the transition probabilities p_{ij} since we will always index the states with natural numbers.

In many physical models the actual state X_n itself cannot be observed or is of less interest. It is thus convenient to consider functions f, $f : S \to \mathbb{R}$ where

[1] We are grateful to Dr. Grodzicki from the Department of Mineralogy, University of Salzburg, Austria, for his aid with this model.

the value $f(\cdot)$ expresses a property of the system which can be measured, e.g. energy, voltage, or speed. Such functions are called "observables". As a natural extension we will also consider observables which are defined on the set of all possible s-tuples of successive states of the chain, where $s \in \mathbb{N}$. In the case of the hydrogen atom, it is especially convenient to work with observables which depend on pairs $(i, j) \in S^2$ of states, where i is the state of departure and j is the destination of the transition which takes place between time points n and $n + 1$. We may, for example, consider the wave length (i.e. energy or "color") of a photon which has been absorbed or emitted. This observable depends on pairs (X_n, X_{n+1}) of successive states.

The theory of Markov chains (MCs) allows a thorough stochastic analysis of such models. In this chapter we shortly recall a classification scheme for MCs which is based on properties of the state space S and the transition probabilities p_{ij}. Spectral theory, when applied to the powers $\mathbb{P}^n$ of the transition matrix $\mathbb{P} = (p_{ij})$, turns out to be a valuable tool in the analysis of asymptotic properties of the process. We also discuss the construction of MCs on the product space S^s, $s \geq 2$, and show that the case of observables which are defined on S^s can be reduced to the case of ordinary observables defined on an "overlapping" chain. We finally quote a main result of the asymptotic theory, the Central Limit Theorem for the average occupancy times of states $i \in S$. We will restrict ourselves to the case of a finite state space in most of the following.

2.1 Markov Chains

The construction of a Markov chain requires two basic ingredients, namely a transition matrix and an initial distribution. We start with the definition of the transition matrix. Assume a finite set $S = \{1, \ldots, m\}$ of states. Assign to each pair $(i, j) \in S^2$ of states a real number p_{ij} such that the properties

$$p_{ij} \geq 0 \quad \forall (i, j) \in S^2 \tag{2.1}$$

$$\sum_{j \in S} p_{ij} = 1 \quad \forall i \in S \tag{2.2}$$

are satisfied and define the transition matrix $\mathbb{P}$ by

$$\mathbb{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & P_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}$$

Let $(X_n)_{n\in\mathbb{N}_0}$ be a sequence of random variables with values in S. Here, n denotes the time at which the state X_n occurs.

Definition 2.1 (Markov Chain) *The sequence $(X_n)_{n\in\mathbb{N}_0}$ is called a homogeneous Markov chain with discrete time, state space S, and transition matrix $\mathbb{P}$, if for every $n \in \mathbb{N}_0$ the condition*

$$P\left[X_{n+1} = j \middle| X_0 = i_0, \ldots, X_n = i_n\right] = P\left[X_{n+1} = j \middle| X_n = i_n\right] = p_{i_n j} \quad (2.3)$$

is satisfied for all $(i_0, \ldots, i_n, j) \in S^{n+2}$, for which $P[X_0 = i_0, \ldots, X_n = i_n] > 0$.

The first identity in (2.3), which is also called "Markov property", defines the "memory" or "order" of the chain. In this case, the order equals one since the transition probabilities are entirely determined by the preceding state. As we will see, the restriction to order-one chains is no serious limitation since processes with arbitrary finite memory s can be interpreted as order-one MCs on the product space S^s. The second identity in (2.3) is called homogeneity condition. It assures that the transition probabilities do not vary with the time n.

So far, we have only specified the ingredients for the evolvement of probabilities throughout the time. To complete the construction of a Markov chain we need to specify an initial distribution. Hence denote by D_S the set of discrete distributions on S,

$$D_S = \{\mathbf{P} = (P_i)_{i\in S} : \ P_i \geq 0, \sum_{i\in S} P_i = 1\},$$

where we represent distributions as row vectors. We call $\mathbf{P}_0 = (P_{0i})_{i\in S} \in D_S$ the *initial distribution* of the chain $(X_n)_{n\in\mathbb{N}_0}$ if $P[X_0 = i] = P_{0i}$ for all states $i \in S$.

Consider now the so-called n-th order transition probabilities $p_{ij}^{(n)} = P[X_{m+n} = j | X_m = i]$, $n \in \mathbb{N}$, which are computed by summing over all

possible intermediate states, that is, over all paths of length $n+1$ with initial state i and final state j,

$$\begin{aligned} p_{ij}^{(n)} &= \sum_{(i_1,\ldots,i_{n-1})\in S^{n-1}} P\Big[X_{m+n}=j\Big|X_m=i, X_{m+1}=i_1,\ldots \\ &\qquad\qquad \ldots, X_{m+n-1}=i_{n-1}\Big] = \\ &= \sum_{(i_1,\ldots,i_{n-1})\in S^{n-1}} p_{ii_1}\cdot\prod_{j=1}^{n-2} p_{i_ji_{j+1}}\cdot p_{i_{n-1}j}, \end{aligned}$$

whence the matrix $(p_{ij}^{(n)})_{(i,j)\in S^2}$ equals the n'th power $\mathbb{P}^n$ of the transition matrix $\mathbb{P}$. We further define $p_{ij}^{(0)} := \delta_{ij}$ so that $\mathbb{P}^0$ equals the identity matrix I_m. Given the distribution $\mathbf{P}_0$ of X_0, the distribution of X_n thus calculates to $\mathbf{P}_0\mathbb{P}^n$, $n\in\mathbb{N}_0$. As we will see, Perron's formula (see Section 2.3) proves to be a powerful tool for the analysis of powers of the matrix $\mathbb{P}$.

The construction of a probability space $(\Omega,\mathcal{A},\mu)$ on which a given chain $(X_n)_{n\in\mathbb{N}_0}$ can be realized is given by the Existence Theorem, see [3, Theorem 8.1] for example. We may think of Ω to consist of all possible paths $(x_0,x_1,x_2,\ldots)$ in $S^{\mathbb{N}_0}$.

The following type of Markov chains deserves special attention.

Definition 2.2 (Independent chain) *Let $\mathbf{P}=(P_1,\ldots,P_m)\in D_S$ and define a Markov chain with state space S, transition matrix $\mathbb{P}=(p_{ij})_{(i,j)\in S^2}$ given by $p_{ij}=P_j$, $i\in S$, and arbitrary initial distribution $\mathbf{P}_0\in D_S$. Then for all $n\in\mathbb{N}_0$,*

$$P\left[X_{n+1}=j\Big|X_0=i_0,\ldots,X_n=i_n\right]=P\left[X_{n+1}=j\Big|X_n=i_n\right]=P_j.$$

We call this the "independent chain" with respect to $\mathbf{P}$.

Corollary 2.1 *In an independent chain the sequence of states $(X_n)_{n\in\mathbb{N}_0}$ is a sequence of independent random variables.*

Independent chains have no memory whence they are also called zero-order Markov chains.

It is quite clear how the above generalizes to the case of countable state space $S=\mathbb{N}$. We shortly recall the following definitions and lemmas which

help to identify the different types of chains. They are discussed in much more detail in every reasonable text book on Markov chains, see e.g. [36], where the reader can also find proofs.

Definition 2.3 (Recurrence and Transience) *Consider the probability*

$$r_{ij} = P\left[\bigcup_{n\in\mathbb{N}} \{X_n = j \big| X_0 = i\}\right]$$

of an eventual visit in state j starting in i. A state i is called recurrent (or persistent) if $r_{ii} = 1$ and transient if $r_{ii} < 1$. A chain is called recurrent (transient) if all states are recurrent (transient).

The following alternative conditions follow by the Borel-Cantelli Lemma:

Lemma 2.2 *A state i is recurrent iff $P[X_n = i$ infinitely often $|X_0 = i] = 1$ or, equivalently, iff $\sum_{n\in\mathbb{N}} p_{ii}^{(n)} = \infty$. On the other hand, it is transient iff $P[X_n = i$ infinitely often $|X_0 = i] = 0$ or, equivalently, iff $\sum_{n\in\mathbb{N}} p_{ii}^{(n)} < \infty$.*

In the case of finite state space S, a state i is transient if and only if it is absorbing, that is, if $p_{ii} = 1$.

Definition 2.4 (Irreducibility) *A MC is called irreducible (or undecomposable) if for all pairs of states $(i, j) \in S^2$ there exists an integer n such that $p_{ij}^{(n)} > 0$.*

Proposition 2.3 *A MC is irreducible if and only if for arbitrary $k \in \mathbb{N}$ and $(i, j) \in S^2$ there exists an integer $n = n_{ij} \geq k$ such that $p_{ij}^{(n)} > 0$.*

Proof: The "if" part being obvious, the "only if" part follows from the fact that for $(i, j) \in S^2$ there exist integers n_{ii}, n_{ij} such that $p_{ii}^{(n_{ii})} > 0$ and $p_{ij}^{(n_{ij})} > 0$ and consequently $p_{ij}^{(ln_{ii}+n_{ij})} \geq (p_{ii}^{(n_{ii})})^l p_{ij}^{(n_{ij})} > 0$ for all $l \in \mathbb{N}$. ■

Irreducible MCs cannot be decomposed into parts which do not interact. It can be shown that either all the states of an irreducible MC are transient or all states are recurrent. In this case, the chain itself is either recurrent or transient. *Finite* irreducible chains are always recurrent, see [3, Example

8.7, p. 118]. Unless indicated otherwise we will assume irreducibility in the following.

In irreducible chains there may still exist a periodic structure such that for each state $i \in S$, the set of possible return times to i when starting in i is a subset of the set $p \cdot \mathbb{N} = \{p, 2p, 3p, \ldots\}$, containing all but a finite set of these elements. The smallest number p with this property is the so-called period of the chain,

$$p = \gcd\{n \in \mathbb{N} \ : \ p_{ii}^{(n)} > 0\}.$$

For $p = 1$, this gives reason for the following definition.

Definition 2.5 (Aperiodicity) *An irreducible chain is called aperiodic (or acyclic) if the period p equals 1 or, equivalently, if for all pairs $(i, j) \in S^2$ of states there is an integer n_{ij} such that for all $n \geq n_{ij}$, the probability $p_{ij}^{(n)} > 0$.*

The state space S of an irreducible periodic chain with period $p > 1$ can be partitioned into p mutually disjoint classes $\mathcal{P}_0, \ldots, \mathcal{P}_{p-1}$ of states such that if the system is in a state $i \in \mathcal{P}_l$ at time n, it will be in a state $j \in \mathcal{P}_k$ at time $n + 1$, where $k \equiv l + 1 \pmod{p}$.

The following lemma gives a class of irreducible and aperiodic chains.

Lemma 2.4 *If $\mathbf{P} = (P_1, \ldots, P_m)$ is strictly positive, $P_i > 0$, $i \in S$, then the independent chain with respect to $\mathbf{P}$ is irreducible and aperiodic.*

Proof: For every $n \in N$, the probability $p_{ij}^{(n)} = P_j$ is positive. ∎

A central role in the analysis of MCs is played by so-called stationary distributions. Such distributions are invariant under the transition matrix.

Definition 2.6 (Stationary distribution) *A distribution $\mathbf{P} \in D_S$ which satisfies $\sum_{i \in S} P_i p_{ij} = P_j$ for all states $j \in S$, or $\mathbf{P}\mathbb{P} = \mathbf{P}$ in matrix notation, is called a stationary distribution.*

By induction we immediately get $\mathbf{P}\mathbb{P}^n = \mathbf{P}$ for all $n \in \mathbb{N}$, so that we have the following lemma:

Lemma 2.5 *If the initial distribution of a chain is a stationary distribution, then the process $(X_n)_{n \in \mathbb{N}_0}$ of states is a stationary process, i.e. each X_n has the same distribution $\mathbf{P}$.*

Lemma 2.6 *For an independent chain with respect to* $\mathbf{P}$, $\mathbf{P}$ *itself is a stationary distribution.*

Proof: $\sum_{i\in S} P_i p_{ij} = \sum_{i\in S} P_i P_j = P_j$ ■

To further analyze stable distributions put

$$T_i = \begin{cases} \infty & : \text{ if } X_n \neq i \quad \forall n \in \mathbb{N} \\ \min\{n \in \mathbb{N} \ : \ X_n = i\} & : \text{ otherwise} \end{cases}$$

the time of the first visit of the chain in state i and let $E_i[T_i]$ denote the expectation conditional on $X_0 = i$, that is, the expected return time to state i.

Definition 2.7 (Positive recurrence) *A state* $i \in S$ *is called positive recurrent (null recurrent) if* $E_i[T_i] < \infty$ *(*$E_i[T_i] = \infty$*). A recurrent chain is called positive recurrent, if all states are positive recurrent.*

Note, that we do not rely on irreducibility in the above definition of a positive recurrent chain.

Lemma 2.7 (Existence and uniqueness of stable distributions) *Positive recurrence of a chain* $(S, \mathbb{P})$ *guarantees the existence of a stable distribution* $\mathbf{P}$. *If, in addition,* $(S, \mathbb{P})$ *is irreducible, then the stable distribution is unique and can be written in the form* $\mathbf{P} = (P_1, P_2, \ldots)$ *with* $P_i = 1/E_i[T_i]$, $i \in S$.

In particular, every *finite* chain is positive recurrent and thus has a stable distribution. For positive recurrent irreducible *aperiodic* chains we have the following limiting result:

Lemma 2.8 *Let* $(S, \mathbb{P})$ *be irreducible, positive recurrent, and aperiodic, with stable distribution* $\mathbf{P}$. *Then* $\mathbf{P} = (P_1, P_2, \ldots)$ *is given by* $P_j = \lim_{n\to\infty} p_{ij}^{(n)}$ *for all pairs* $(i, j) \in S^2$ *of states.*

In this case, $\mathbb{P}^n$ converges element-wise to the transition matrix of the independent chain with respect to the stable distribution $\mathbf{P}$. For the corresponding l_1-result on the sequence of distributions $\mathbf{P}_0\mathbb{P}^n$, $\mathbf{P}_0 \in D_S$ an arbitrary initial distribution, see [6, Theorem 6.1] and [42]. If S is finite in addition, we have the following Lemma on the speed of convergence:

Lemma 2.9 *Let $(S, \mathbb{P})$ be a finite irreducible aperiodic MC with stationary distribution* $\mathbf{P}$*. Then there exist $c \geq 0$ and $0 \leq \rho < 1$, such that*

$$|p_{ij}^{(n)} - P_j| \leq c\rho^n, \quad \forall (i,j) \in S^2, \forall n \in \mathbb{N}.$$

Any arbitrary initial distribution converges exponentially to $\mathbf{P}$ in this case. This class of chains allows a particular simple form of the Central Limit Theorem for the number of visits in each state. For the corresponding result (see Theorem 2.18) both, the finiteness and irreducibility are essential prerequisites. Although aperiodicity may be omitted, we will have this property in the examples discussed in Chapter 3.

2.2 Coupled-, Overlapping- and Higher-Order Chains

The two construction schemes for Markov chains which we introduce in this section will prove fruitful in Section 3.1. So-called coupled chains describe the joint behavior of independent copies of a chain where each copy evolves according to the laws of the original chain. Overlapping chains on the other hand arise from considering successive overlapping tuples of states $(X_n, X_{n+1}, \ldots, X_{n+s-1})$. They can be employed to analyze stochastic properties of transitions in the original chain. The section is finished with some remarks on chains of order higher than one, that is, chains which have an arbitrary finite memory. We will see that nothing new is involved since we can represent such a chain by an ordinary chain of order one. For the reason of simplicity we will assume finiteness of S in most of the proofs although the constructions apply also to the case of countable S. With respect to the Central Limit Theorem, Theorem 2.18, we will analyze conditions for irreducibility and aperiodicity. In the following, denote by S^s the s-fold Cartesian product of S, denote its elements by $\mathbf{i} = (i_1, \ldots, i_s) \in S^s$ and let S^s be ordered lexicographically.

2.2.1 Coupled Chains

Consider a chain $(S, \mathbb{P})$ and the according sequence $(X_n)_{n \in \mathbb{N}_0}$ of random variables. Assume s independent copies $(X_n^{(1)})_{n \in \mathbb{N}_0}, \ldots, (X_n^{(s)})_{n \in \mathbb{N}_0}$

where each chain $(X_n^{(i)})_{n\in\mathbb{N}_0}$, $X_n^{(i)} \in S$, evolves according to its own initial distribution $\mathbf{P}_0^{(i)}$ and according to the common transition matrix $\mathbb{P}$. As an example consider s identical games being played independently but in a synchronized manner. We are interested in the vectors $\overline{X}_n \in S^s$, $\overline{X}_n = (X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(s)})$ which describe the state of all chains (i.e. games) at each time instant $n \in \mathbb{N}_0$. This sequence can again be modelled as a chain with state space S^s and transition matrix $\overline{\mathbb{P}}$ containing the probabilities to go from a state $(X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(s)}) = (i_1, \ldots, i_s) = \mathbf{i}$ to a state $(X_{n+1}^{(1)}, X_{n+1}^{(2)}, \ldots, X_{n+1}^{(s)}) = (j_1, \ldots, j_s) = \mathbf{j}$.

Definition 2.8 (Coupled Chain) *Let $(S, \mathbb{P})$ be a Markov chain and let $s \geq 2$ be an integer. The chain $(\tilde{S}, \tilde{\mathbb{P}})$ with state space $\tilde{S} = S^s$ is called a coupled chain with dimension s, if $\tilde{\mathbb{P}} = (\tilde{p}_{\mathbf{ij}})$, $(\mathbf{i}, \mathbf{j}) \in \tilde{S}^2$, is given by $\tilde{p}_{\mathbf{ij}} = \prod_{k=1}^{s} p_{i_k j_k}$. Further let $\Psi : D_S \to D_{\tilde{S}}$ be defined by $\Psi(\mathbf{P}) = \tilde{\mathbf{P}} = (\tilde{P}_{\mathbf{i}})_{\mathbf{i}\in\tilde{S}}$, where $\tilde{P}_{\mathbf{i}} = \prod_{k=1}^{s} P_{i_k}$ for $\mathbf{P} = (P_i)_{i\in S}$.*

It is easy to check conditions (2.1) and (2.2) on $\tilde{\mathbb{P}}$ and to prove that $\tilde{\mathbf{P}} \in D_{\tilde{S}}$. The coupled chain trivially fulfills the Markov property and is homogeneous. The components of the random vector $\overline{X}_n$ are independent and the projection of the sequence $(\overline{X}_n)_{n\in\mathbb{N}}$ to the k'th component $(\overline{X}_n^{(k)})_{n\in\mathbb{N}}$ is the Markov chain $(S, \mathbb{P})$ with initial distribution $\mathbf{P}_0^{(k)}$.

Lemma 2.10 *If $(S, \mathbb{P})$ is an independent chain with respect to $\mathbf{P}$, then the coupled chain with dimension s is independent with respect to $\Psi(\mathbf{P})$.*

The Lemma follows directly from Definition 2.8. The next Lemma describes two important properties of coupled chains.

Lemma 2.11 *Let $(\tilde{S}, \tilde{\mathbb{P}})$ be the coupled chain with dimension s of $(S, \mathbb{P})$.*
(i) A necessary and sufficient condition for the irreducibility of $(\tilde{S}, \tilde{\mathbb{P}})$ is that $(S, \mathbb{P})$ is irreducible and aperiodic. $(\tilde{S}, \tilde{\mathbb{P}})$ is aperiodic in this case, too.
(ii) If $\mathbf{P}$ is a stationary distribution of $(S, \mathbb{P})$, then $\Psi(\mathbf{P})$ is a stationary distribution of the coupled chain.

Proof: As to (i), we first show sufficiency. Since $(\tilde{\mathbb{P}})^n$ equals $\tilde{\mathbb{P}^n}$, we have for each $(\mathbf{i}, \mathbf{j}) \in \tilde{S}^2$, $\tilde{p}_{\mathbf{ij}}^{(n)} = \prod_{k=1}^{s} p_{i_k j_k}^{(n)}$ which is positive for all $n \in \mathbb{N}$ exceeding a certain $n_{\mathbf{i},\mathbf{j}}$ if $(S, \mathbb{P})$ is irreducible and aperiodic. Necessity of the irreducibility of $(S, \mathbb{P})$ is trivial since if for all $(\mathbf{i}, \mathbf{j}) \in \tilde{S}^2$ there exists

a $n \in \mathbb{N}$ such that $\tilde{p}_{\mathbf{ij}}^{(n)} > 0$ then for all $k \in \{1, \ldots, s\}$ we have $p_{i_k j_k}^{(n)} > 0$ so that the irreducibility of $(S, \mathbb{P})$ follows from that of $(\tilde{S}, \tilde{\mathbb{P}})$. Now for the necessity of the aperiodicity of $(S, \mathbb{P})$. If $(S, \mathbb{P})$ has period $p > 1$ we can always find two states $(\mathbf{i}, \mathbf{j}) \in \tilde{S}^2$ such that $i_k = j_k$ for all $k \in \{1, \ldots, s-1\}$ and i_s and j_s belong to different periodic classes of the chain $(S, \mathbb{P})$. In this case, the probabilities $p_{i_k j_k}^{(n)}$, $1 \leq k \leq s-1$, are positive if and only if p divides n whereas $p_{i_s j_s}^{(n)} = 0$ in that case such that $\tilde{p}_{\mathbf{ij}}^{(n)} = 0$ for all $n \in \mathbb{N}$. The proof for (ii) is given by

$$\sum_{\mathbf{i} \in \tilde{S}} \tilde{P}_{\mathbf{i}} \tilde{p}_{\mathbf{ij}} = \prod_{k=1}^{s} \left(\sum_{i \in S} P_i p_{i j_k} \right) = \prod_{k=1}^{s} P_{j_k} = \tilde{P}_{\mathbf{j}},$$

where the second equality follows from the fact that $\mathbf{P}$ was assumed to be a stationary distribution of $(S, \mathbb{P})$. ∎

As a simple example, where the original chain is irreducible and periodic with period $p = 2$ and the coupled chain for $s = 2$ is not irreducible, consider the chain $(S, \mathbb{P})$ with $S = \{1, 2\}$ and $\mathbb{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, such that in $(\tilde{S}, \tilde{\mathbb{P}})$ any path of finite length from the state $(1, 1)$ to the state $(1, 2)$ has zero probability. Coupled chains can be used as a technical tool in the derivation of properties of (ordinary) Markov chains, see e.g. [3, Theorem 8.6.]. More concrete examples are given in Section 3.1.

2.2.2 Overlapping Chains

In the introduction to this Chapter we mentioned the importance of observables which depend on the transition which takes place from time n to $n+1$. In this case, it is useful to consider the sequence of pairs $(X_n, X_{n+1})_{n \in \mathbb{N}_0}$ of successive states. Given a Markov chain $(X_n)_{n \in \mathbb{N}_0}$ with state space S and transition matrix $\mathbb{P}$ we will more generally analyze the sequence of overlapping s-tuples $(X_n, X_{n+1}, \ldots, X_{n+s-1})$ with transitions from $(X_n, X_{n+1}, \ldots, X_{n+s-1}) = (i_1, \ldots, i_s)$ to $(X_{n+1}, X_{n+2}, \ldots, X_{n+s}) = (j_1, \ldots, j_s)$. This turns out to be a Markov chain itself which we call the overlapping chain with dimension s:

Definition 2.9 (Overlapping Chain) *Let $(S, \mathbb{P})$ be a Markov chain and*

let $s \geq 2$ be an integer. The chain $(\overline{S}, \overline{\mathbb{P}})$ with state space $\overline{S} \subset S^s$,

$$\overline{S} = \left\{ (i_1, \ldots, i_s) \in S^s \ : \ \prod_{k=1}^{s-1} p_{i_k i_{k+1}} > 0 \right\}$$

is called overlapping chain with dimension s if $\overline{\mathbb{P}} = (\overline{p}_{\mathbf{ij}})$, $(\mathbf{i}, \mathbf{j}) \in \overline{S}^2$ is given by

$$\overline{p}_{\mathbf{ij}} = \begin{cases} p_{i_s j_s} & \text{if } i_{k+1} = j_k \text{ for every } k \in \{1, \ldots, s-1\} \\ 0 & \text{otherwise.} \end{cases}$$

Further let $\Upsilon : D_S \to D_{\overline{S}}$ be defined by $\Upsilon(\mathbf{P}) = \overline{\mathbf{P}} = (\overline{P_{\mathbf{i}}})_{\mathbf{i} \in \overline{S}}$, where $\overline{P_{\mathbf{i}}} = P_{i_1} \cdot \prod_{k=1}^{s-1} p_{i_k i_{k+1}}$ for $\mathbf{P} = (P_i)_{i \in S}$. Here we use the lexicographical order on S^s for the components of the vector.

It is again easy to check conditions (2.1) and (2.2) on $\overline{\mathbb{P}}$, to assert the Markov property and homogeneity, and to prove that $\overline{\mathbf{P}} \in D_{\overline{S}}$.

Lemma 2.12 *If $(X_n)_{n \in \mathbb{N}_0}$ is a MC with state space S, transition matrix $\mathbb{P}$ and initial distribution $\mathbf{P}_0$, then the sequence $(\overline{X}_n)_{n \in \mathbb{N}}$, $\overline{X}_n = (X_n, X_{n+1}, \ldots, X_{n+s-1})$, is the overlapping chain with dimension s, state space $\overline{S}$, transition matrix $\overline{\mathbb{P}}$, and initial distribution $\overline{\mathbf{P}}_0 = \Upsilon(\mathbf{P}_0)$.*

Proof: Let $(i_0, i_1, \ldots, i_{l+s-1}) \in S^{s+l}$ such that $\mathbf{i}_k = (i_k, \ldots, i_{k+s-1})$, $k \in \{0, \ldots, l\}$, are states of the overlapping chain (i.e. elements of $\overline{S}$). Then

$$P[\overline{X}_0 = \mathbf{i}_0, \ldots, \overline{X}_l = \mathbf{i}_l] = P_{0 i_0} \prod_{k=0}^{l+s-1} p_{i_k i_{k+1}} = \overline{\mathbf{P}}_{0 \mathbf{i}_0} \prod_{k=0}^{l-1} \overline{p}_{\mathbf{i}_k \mathbf{i}_{k+1}}.$$

Otherwise, the sequence $\overline{X}_0 = \mathbf{i}_0, \ldots, \overline{X}_l = \mathbf{i}_l$ can, in fact, not occur and the according probability in the chain $(\overline{S}, \overline{\mathbb{P}})$ is consistently zero by definition. ■

Lemma 2.13 *Let $(\overline{S}, \overline{\mathbb{P}})$ be the overlapping chain with dimension s of $(S, \mathbb{P})$.*
(i) $(\overline{S}, \overline{\mathbb{P}})$ is irreducible if and only if $(S, \mathbb{P})$ is irreducible.
(ii) $(\overline{S}, \overline{\mathbb{P}})$ is aperiodic if and only if $(S, \mathbb{P})$ is aperiodic.
(iii) If $\mathbf{P}$ is a stationary distribution of $(S, \mathbb{P})$, then $\Upsilon(\mathbf{P})$ is a stationary distribution of the overlapping chain.

Proof: It suffices to consider n'th order transition probabilities $\overline{p}_{\mathbf{ij}}^{(n)}$, $(\mathbf{i},\mathbf{j}) \in \overline{S}^2$, with $n \geq s+1$, such that the tuples

$$\overline{X}_m = (X_m, \ldots, X_{m+s-1}) = (i_1, \ldots, i_s) = \mathbf{i}$$

and

$$\overline{X}_{m+n} = (X_{m+n}, \ldots, X_{m+n+s-1}) = (j_1, \ldots, j_s) = \mathbf{j}$$

of the original chain $(S, \mathbb{P})$ do not overlap and there is at least one intermediate state X_{m+s} between the occurrence of i_s and j_1. Denote the intermediate states by $u_1, \ldots, u_{n-s}$. For every $(\mathbf{i},\mathbf{j}) \in \overline{S}^2$ the probability $\overline{p}_{\mathbf{ij}}^{(n)}$ is equal to the probability to jump from state i_s to the state j_1 in $n-s+1$ steps and to proceed by visiting the states $j_2, \ldots, j_s$ thereafter. Since paths with zero probability in $(S, \mathbb{P})$ are exactly those which contain s-tuples of states which are not contained in $\overline{S}$, we have for every $m \in \mathbb{N}_0$,

$$\begin{aligned}
\overline{p}_{\mathbf{ij}}^{(n)} &= \sum_{(u_1,\ldots,u_{n-s}) \in S^{n-s}} P\Big[\overline{X}_{m+n} = \mathbf{j} \Big| X_{m+s-1} = i_s, X_{m+s} = u_1, \ldots \\
&\qquad\qquad \ldots, X_{m+n-1} = u_{n-s}\Big] = \\
&= p_{i_s j_1}^{(n-s+1)} \prod_{k=1}^{s-1} p_{j_k j_{k+1}},
\end{aligned}$$

where $\prod_{k=1}^{s-1} p_{j_k j_{k+1}} > 0$ since $\mathbf{j} \in \overline{S}$. Thus $\overline{p}_{\mathbf{ij}}^{(n)} > 0$ if and only if $p_{i_s j_1}^{(n-s+1)} > 0$. Now (i) and (ii) follow since for every $(i,j) \in S^2$ there exists a pair $(\mathbf{i},\mathbf{j}) \in \overline{S}^2$ such that $i = i_s$ and $j = j_1$ on the one hand and – trivially – for each $(\mathbf{i},\mathbf{j}) \in \overline{S}^2$ we have that $(i_s, j_1) \in S^2$.
To prove (iii) note that $\overline{p}_{\mathbf{ij}} > 0$ implies that $i_{k+1} = j_k$ for every $1 \leq k \leq s-1$. Thus

$$\begin{aligned}
\sum_{\mathbf{i} \in \overline{S}} \overline{P_{\mathbf{i}}} \overline{p}_{\mathbf{ij}} &= \sum_{\mathbf{i} \in \overline{S}} P_{i_1} \prod_{k=1}^{s-1} p_{i_k i_{k+1}} p_{i_s j_s} \mathbf{1}_{\{i_k = j_{k+1} : k \in \{1,\ldots,s-1\}\}} = \\
&= P_{i_1} p_{i_1 j_1} \prod_{k=1}^{s-1} p_{j_k j_{k+1}} = P_{j_1} \prod_{k=1}^{s-1} p_{j_k j_{k+1}} = \overline{P_{\mathbf{j}}}
\end{aligned}$$

■

Examples of overlapping chains are discussed in Section 3.1.

2.2.3 Higher Order Chains

We finish this section by a construction which is very similar to that of overlapping chains and allows the analysis of higher order chains within the framework of ordinary chains. Recall the concept of a homogenous Markov chain of order s: We assume a finite state space S and a matrix $\mathbb{P} = (p_{\mathbf{i}j})_{(\mathbf{i},j)\in S^s\times S}$ of real nonnegative values satisfying $\sum_{j\in S} p_{\mathbf{i}j} = 1$ for every $\mathbf{i} \in S^s$. Here, as usual, we assume lexicographical ordering on S^s. A sequence of random variables $(X_n)_{n\in\mathbb{N}_0}$, $X_n \in S$, is called Markov chain of order s with transition matrix $\mathbb{P}$ if for arbitrary $n \geq s-1$ and $(u_0,\ldots,u_n) \in S^{n+1}$, $j \in S$, with $P[X_0 = u_0,\ldots,X_n = u_n] > 0$, where we put $\mathbf{i} = (i_1,\ldots,i_s) = (u_{n-s+1},\ldots,u_n)$, the condition

$$\begin{aligned} P\left[X_{n+1} = j \middle| X_0 = u_0,\ldots,X_n = u_n\right] &= \\ P\left[X_{n+1} = j \middle| X_{n-s+1} = i_1,\ldots,X_n = i_s\right] &= p_{\mathbf{i}j} \end{aligned} \tag{2.4}$$

is satisfied.

Actually, nothing new is involved here in comparison to ordinary chains. To see this, consider the following generalization of the notion of an overlapping chain with dimension s of an ordinary order-one chain to the case of chains with order s. Define the state space

$$\overline{S} = \{(i_2,\ldots,i_s,j) \in S^s \ : \ \exists i_1 \in S \text{ such, that for } \mathbf{i} = (i_1,\ldots,i_s) \ : \ p_{\mathbf{i}j} > 0\}$$

and the transition matrix $\overline{\mathbb{P}} = (\overline{p}_{\mathbf{ij}})$, $(\mathbf{i},\mathbf{j}) \in \overline{S}^2$,

$$\overline{p}_{\mathbf{ij}} = \begin{cases} p_{\mathbf{i}j_s} & \text{if } i_{k+1} = j_k \text{ for every } k \in \{1,\ldots,s-1\} \\ 0 & \text{otherwise.} \end{cases}$$

$(\overline{S},\overline{\mathbb{P}})$ is a Markov chain of order one since – owing to (2.4) – the probability of a state $\overline{X}_{n+1}$ is entirely determined by its predecessor $\overline{X}_n$. Chains of arbitrary but fixed order allow to describe a huge range of models which occur, for example, in the field of linear control with noise, see [36, Sections 1.2.1 and 1.2.2].

2.3 Spectral Analysis of Markov Chains

By (2.1) and (2.2), the matrix $\mathbb{P}$ of transition probabilities of a Markov chain represents a particular species of non-negative matrices. We call such

matrices stochastic (in the wide sense). If, in addition, we have that at least one element p_{ij} in each column is greater than zero,

$$\forall j \in S,\ \exists i \in S \text{ such that } p_{ij} > 0, \tag{2.5}$$

we call $\mathbb{P}$ a stochastic matrix in the restricted sense[2]. For finite chains (2.5) is equivalent to the fact that there are no transient states which is, for example, true for irreducible chains. In this section we will only consider finite chains and we always let the state space be $S = \mathbb{N}_m$.

The theory of stochastic matrices is strongly built upon the analysis of the eigenvalues $\lambda_k \in \mathbb{C}$ of $\mathbb{P}$. It is clear from (2.1) and (2.2) that, first, there exists an eigenvalue $\lambda_0 = 1$ with right-eigenvector $(1, \ldots, 1)'$, and that, secondly, the spectral radius $\rho(\mathbb{P}) = \sup\{|\lambda| : |\lambda I_m - \mathbb{P}| = 0\}$ equals 1 since it can be bounded from below and above by

$$1 = \min_{i \in S} \sum_{j \in S} p_{ij} \leq \rho(\mathbb{P}) \leq \max_{i \in S} \sum_{j \in S} p_{ij} = 1.$$

For every stochastic matrix we thus have $\rho(\mathbb{P}) = 1$. All other eigenvalues λ_k, $k \neq 0$, are smaller than or equal to 1 in absolute value. Note, that the stable distribution $\mathbf{P}$ is a left-eigenvector to the eigenvalue $\lambda_0 = 1$. The following Theorem summarizes some facts about the eigenvalues of $\mathbb{P}$.

Theorem 2.14 *If $(S, \mathbb{P})$ is an irreducible Markov chain with finite state space S, then the multiplicity of the eigenvalue $\lambda_0 = 1$ equals 1. If $(S, \mathbb{P})$ is aperiodic, in addition, then the modulus of all other eigenvalues λ_k, $k \neq 0$ of $\mathbb{P}$ is smaller then 1, $|\lambda_k| < 1$ for $k \geq 1$. If, on the other hand, $(S, \mathbb{P})$ is periodic with period $p > 1$, then all p'th roots of unity (namely $e^{i2\pi j/p} \in \mathbb{C}$, $j \in \{0, \ldots, p-1\}$) are eigenvalues of $\mathbb{P}$, each with multiplicity 1, and there are no other eigenvalues with modulus 1*

We refer the reader to [23, Theorems 2.1, 3.1, 3.2] for a proof of Theorem 2.14, and to the first chapter in [50] for further details on the analysis of eigenvalues and eigenvectors of non-negative matrices. In the following we express the stable distribution of an irreducible finite chain in terms of the matrix $\mathbb{P}$. In addition we present a limiting result which follows by the aid of the so-called Perron formula for powers of square matrices. A similar

[2]Note that in [49] the term stochastic matrix is used for stochastic matrices in the restricted sense.

technique applies to the characteristic function of the frequencies in a chain and leads to a form of the Central Limit Theorem which is quite accessible to numerical computations. For details we refer the reader to the book of Romanovsky, [49, Chapters 1 and 4].

2.3.1 Perron's Formula

Perron's Formula expresses the elements $a_{ij}^{(n)}$ of the n-th power A^n of a $m \times m$ square matrix A in terms of the eigenvalues and minor determinants of A. An eigenvalue λ_k with multiplicity ρ_k of A is a root with multiplicity ρ_k of the characteristic polynomial[3] $A(\lambda) = |\lambda I_m - A|$, where I_m is the identity matrix. Denote the minor of a matrix A which arises from deleting the i-th row and j-th column by A_{ij} and denote the so-called minor of the determinant $A(\lambda)$ which equals $(-1)^{i+j}$ times the determinant of the minor of $(\lambda I_m - A)_{ji}$ by $A_{ij}(\lambda)$. Note the correct order of i and j in the latter definition. From [49, p. 16] we cite the following theorem:

Theorem 2.15 (Perron's Formula) *Let $\lambda_0, \lambda_1, \ldots, \lambda_\mu$ be the eigenvalues of a $m \times m$ matrix A with multiplicities $\rho_0, \ldots, \rho_\mu$, respectively. Define for every $k \in \{0, \ldots, \mu\}$ a polynomial π_k of degree $m - \rho_k$ by $A(\lambda) = (\lambda - \lambda_k)^{\rho_k} \pi_k(\lambda)$. Then for all $(i,j) \in S^2 = \mathbb{N}_m^2$, and for all $n \in \mathbb{N}_0$,*

$$a_{ij}^{(n)} = \sum_{k=0}^{\mu} \frac{1}{(\rho_k - 1)!} \frac{\partial^{\rho_k - 1}}{\partial \lambda} \left[\frac{\lambda^n A_{ij}(\lambda)}{\pi_k(\lambda)} \right]_{\lambda = \lambda_k} .$$

As we will see below, Perron's formula is particularly useful if the matrix A has a maximal eigenvalue $\lambda_0 = 1$ with multiplicity 1 and if all other eigenvalues satisfy $|\lambda_k| < 1$. This justifies the following definition and lemma which we cite from [49, pp. 21 and 46].

Definition 2.10 *A chain $(S, \mathbb{P})$ is called regular if $\mathbb{P}$ has a maximal eigenvalue $\lambda_0 = 1$ with multiplicity 1 and all other eigenvalues satisfy $|\lambda_k| < 1$. The chain is called positively regular if, in addition, all the numbers P_j are strictly positive.*

Lemma 2.16 *A chain is regular (positively regular) if and only if there exists an integer n such that for at least one $j \in S$ (for all $j \in S$) and for all*

[3] Here we follow the notation in [49, p. 5].

$i \in S$ the probability $p_{ij}^{(n)} > 0$. If a regular chain is irreducible, it is positively regular.

For a proof see [49, Theorem 14.I and 14.II]. The next corollary, which follows trivially from the definitions of irreducibility and aperiodicity, gives an important class of positively regular chains.

Corollary 2.17 *Irreducible aperiodic chains are positively regular.*

By Corollary 2.17 the transition matrix $\mathbb{P}$ of an irreducible aperiodic chain has an eigenvalue $\lambda_0 = 1$ with multiplicity 1 and satisfies $|\lambda_i| < \lambda_0$ for all $1 \leq i \leq \mu$. In this case, Perron's formula for the n'th power of $\mathbb{P}$ assumes the following form, see [49, p. 19]:

$$p_{ij}^{(n)} = P_j + \sum_{k=1}^{\mu} \frac{1}{(\rho_k - 1)!} \frac{\partial^{\rho_k - 1}}{\partial \lambda} \left[\frac{\lambda^n \mathbb{P}_{ij}(\lambda)}{\pi_k(\lambda)} \right]_{\lambda = \lambda_k}, \tag{2.6}$$

where

$$P_j = \frac{\mathbb{P}_{jj}(1)}{\sum_{i=1}^{m} \mathbb{P}_{ii}(1)}. \tag{2.7}$$

Here $p_{ij}^{(n)} \to P_j$ as $n \to \infty$ since the remaining term

$$\sum_{k=1}^{\mu} \frac{1}{(\rho_k - 1)!} \frac{\partial^{\rho_k - 1}}{\partial \lambda} \left[\frac{\lambda^n \mathbb{P}_{ij}(\lambda)}{\pi_k(\lambda)} \right]_{\lambda = \lambda_k} \tag{2.8}$$

tends to zero on behalf of $|\lambda_k| < 1$, $k \in \{1, \ldots, \mu\}$, see [49, p. 21].

Since the irreducible aperiodic chain $(S, \mathbb{P})$ has a finite state space here, we already know from Lemma 2.8 that for all pairs $(i, j) \in S^2$ of states the n'th order transition probabilities $p_{ij}^{(n)}$ tend to the probabilities P_j of the stable distribution $\mathbf{P} = (P_1, \ldots, P_m)$, i.e. it holds $\lim_{n \to \infty} p_{ij}^{(n)} = P_j$. By Lemma 2.7, the stable distribution is representable as $P_i = 1/E_i[T_i]$. Thus we get the following two representations for the stable distribution $\mathbf{P} = (P_1, \ldots, P_m)$:

$$P_j = 1/E_j[T_j] = \frac{\mathbb{P}_{jj}(1)}{\sum_{i=1}^{m} \mathbb{P}_{ii}(1)}, \quad i \in S. \tag{2.9}$$

Moreover, in [49, Chapter 3] it is shown that (2.7) also represents the stable distribution in the case of a periodic irreducible chain although the remaining term (2.8) will not tend to zero in this case due to the existence of eigenvalues $\lambda_k \neq 1$ with $|\lambda_k| = 1$, recall Theorem 2.14.

2.3.2 Frequency Analysis

As before let $S = \{1, \ldots, m\}$ and $\mathbb{P} = (p_{ij})_{(i,j)\in S^2}$ and denote the state of the chain $(S, \mathbb{P})$ at time l by X_l, $l \in \mathbb{N}_0$. For $n \in \mathbb{N}$ we consider the occupation times $C_i^{(n)} := \#\{l : X_l = i, 0 \leq l \leq n-1\}$ of each state $i \in S$ in the time interval $l \in \{0, \ldots, n-1\}$ and define the vector $C^{(n)} = (C_1^{(n)}, \ldots, C_m^{(n)})$. One might expect both, a strong law of large numbers and a Central Limit Theorem (CLT) for a suitable normalized sequence $(C^{(n)})_{n\in\mathbb{N}}$. In the following we only discuss the CLT for irreducible aperiodic chains. We consider the characteristic function of the vector $C^{(n)}$ and deduce its first and second central moments. From these, the proper normalization sequence for the CLT, which we cite from [49], follows readily. For details on the strong law of large numbers for Markov chains, the reader is referred to the exhaustive monograph of Meyn and Tweedie, [36, Theorem 13.0.1], and to [49, p. 338].

The characteristic function $\theta^{(n)}$ of the random vector $C^{(n)}$ at the point $\mathbf{t} = (t_1, \ldots, t_m) \in \mathbb{R}^m$ is the value $\theta^{(n)}(\mathbf{t}) = E[\exp(i\mathbf{t} \cdot C^{(n)})]$, where $i = \sqrt{-1}$, and $\cdot$ denotes the inner product. $\theta^{(n)}$ can be expressed in the following form which we cite from [49, Chapter 4, Number 37]: define the $m \times m$ matrix $U = (u_{ij})_{(i,j)\in S^2}$ by $u_{ij} = p_{ij}\exp(it_j)$ and the vector $U_0 = (u_{01}, \ldots, u_{0m})$ by $u_{0i} = p_{0i}\exp(it_i)$, where the p_{ij}, $(i,j) \in S^2$ are the transition probabilities, and p_{0i} are the probabilities of the initial distribution $\mathbf{P}_0$ of the chain. Then

$$\theta^{(n)}(\mathbf{t}) = \sum_{i\in S}\sum_{j\in S} u_{0i}u_{ij}^{(n)}, \tag{2.10}$$

where $u_{ij}^{(n)}$ denotes the element in row i and column j of the n-th power U^n of the matrix U. This representation allows the use of Perron's formula.

Every moment of finite order can be calculated by differentiating (2.10) and evaluating at the point $\mathbf{t} = \emptyset$. For the case of irreducible chains, Romanovsky deduces the following expressions for the asymptotic mean $\mathbf{E} := \lim_{n\to\infty} \frac{1}{n}E[C^{(n)}]$ and covariance matrix $V = (v_{ij})_{(i,j)\in S^2} := \lim_{n\to\infty} \frac{1}{n}V[C^{(n)}]$: by (2.9) the stable distribution $\mathbf{P} = (P_1, \ldots, P_m)$ of the irreducible chain $(S, \mathbb{P})$ is given by

$$P_i = \frac{\mathbb{P}_{ii}(1)}{\sum_{k=1}^m \mathbb{P}_{kk}(1)}, \quad i \in S.$$

The asymptotic mean $\mathbf{E}$ equals this stable distribution,

$$\mathbf{E} = \lim_{n\to\infty} \frac{1}{n} E[C^{(n)}] = \mathbf{P}. \tag{2.11}$$

For the case of an aperiodic chain the proof follows from the fact, that the associated sequence of Cesaro means of the convergent sequence $(p_{ij}^{(n)})_{n\in\mathbb{N}_0}$ with limit $\lim_{n\to\infty} p_{ij}^{(n)} = P_j$ converges to the same limit P_j so that

$$\frac{1}{n} E[C_j^{(n)}] = \sum_{i\in S} P[X_0 = i] \frac{1}{n} \sum_{l=0}^{n-1} p_{ij}^{(n)} \to \sum_{i\in S} P[X_0 = i] P_j = P_j.$$

The general case of an irreducible finite chain is treated in [49, Chapter 4, (38.14)].

As to V, let us denote the minor of second order of a matrix A which arises from deleting the i'th and j'th row and the k'th and l'th column by $A_{ij|kl}$. If $i = k$ and $j = l$ we call $A_{ij|ij}$ a principal minor of second order of the matrix A. Denote the so-called principal minor of second order of the determinant $A(\lambda)$, namely the determinant of the principal minor of $(\lambda I_m - A)_{ij|ij}$, by $A_{ij|ij}(\lambda)$. For arbitrary $(i,j) \in S^2$ put

$$Q_{ij} = \frac{\mathbb{P}_{ij|ij}(1)}{\sum_{k\in S} \mathbb{P}_{kk}(1)}, \quad Q_i = \sum_{j\neq i} Q_{ij}, \text{ and } Q = \sum_{i\in S} Q_i.$$

According to [49, Chapter 4, p. 182] the asymptotic covariance matrix $V = (v_{ij})_{(i,j)\in S^2}$ is given by

$$\begin{aligned}
v_{ii} &= P_i^2(1-Q) + 2P_iQ_i - P_i = && (2.12)\\
&= -{P_i}^2 + 2P_i(P_i(1-\frac{Q}{2}) + Q_i) - P_i = && (2.13)\\
&= P_i(1-P_i) + 2P_i(P_i(1-\frac{Q}{2}) + Q_i - 1), \quad i \in S && (2.14)\\
v_{ij} &= P_iP_j(1-Q) + P_iQ_j + P_jQ_i - Q_{ij} = && (2.15)\\
&= -P_iP_j + P_i(P_j(1-\frac{Q}{2}) + Q_j) + && (2.16)\\
&\quad +P_j(P_i(1-\frac{Q}{2}) + Q_i) - Q_{ij}, \quad (i,j) \in S^2, i \neq j.
\end{aligned}$$

In (2.14) and (2.16), the similarity to the multinomial distribution becomes evident. If the vector $C^{(n)}$ was distributed multinomial with parameters

n and $\mathbf{P}$, $C^{(n)} \sim \mathcal{MN}(n, \mathbf{P})$, then for all $n \in \mathbb{N}$, $\frac{1}{n}E[C^{(n)}] = \mathbf{P}$, and $\frac{1}{n}V[C^{(n)}] = V$ with $V = (v_{ij})_{(i,j)\in S^2}$, $v_{ii} = P_i(1 - P_i)$, and $v_{ij} = -P_iP_j$ for $j \neq i$. The additional terms in (2.14) and (2.16) are corrections to the variances and covariances which arise from the correlation within the chain. As we will see in Section 3.1.1, these terms consistently are zero for an independent chain, which is a model for a multinomial distributed vector $C^{(n)}$.

An important role is played by the rank $r = R(V)$ of the asymptotic covariance matrix V. It is well known that the rank of the covariance matrix of a multinomial distribution with parameters n and $\mathbf{P} = (P_1, \ldots, P_m)$, where $P_i > 0$, $i \in \{1, \ldots, m\}$, equals $m-1$. Thus the asymptotic covariance matrix V of the counter vector $C^{(n)}$ of an independent chain $(S, \mathbb{P})$ with respect to a strictly positive vector $\mathbf{P}$ has rank $m - 1$. Trivially, $r \leq m - 1$ for *every* chain since the sum $\sum_{i\in S} C_i^{(n)} = n$ so that the counter vector lives on a linear subspace of $\mathbb{N}^m$ with dimension less or equal $m - 1$.

Since the coupled chain $(\tilde{S}, \tilde{\mathbb{P}})$ with dimension $s > 1$ of $(S, \mathbb{P})$ is independent with respect to $\Psi(\mathbf{P})$ by Lemma 2.10, and $\Psi(\mathbf{P})$ is strictly positive in this case, the rank $R(\tilde{V})$ of the asymptotic covariance matrix $\tilde{V}$ of $(\tilde{S}, \tilde{\mathbb{P}})$ equals $m^s - 1$. The case of an overlapping chain $(\overline{S}, \overline{\mathbb{P}})$ with dimension $s > 1$ of the independent chain $(S, \mathbb{P})$ with respect to $\mathbf{P}$ is more complicated. As shown in [55, p. 115], the rank $R(\overline{V})$ of the asymptotic covariance matrix $\overline{V}$ of the chain $(\overline{S}, \overline{\mathbb{P}})$ equals $R(\overline{V}) = m^s - m^{s-1}$.

2.3.3 The Central Limit Theorem

As before, let $(S, \mathbb{P})$ be a chain with state space $S = \{1, \ldots, m\}$ and transition matrix $\mathbb{P}$. We again consider the counter vector $C^{(n)} = (C_1^{(n)}, \ldots, C_m^{(n)}) \in \mathbb{N}_0^m$ counting the number of visits of the chain in each state $i \in S$. From [49, Theorem 42.VII] we cite the following Central Limit Theorem, which we present in a slightly modified version.

Theorem 2.18 *Let $(S, \mathbb{P})$ be an irreducible, aperiodic, and finite chain and let $\mathbf{P} = \mathbf{E}$ and V be defined as above, then*

$$\sqrt{n}\left(\frac{1}{n}C^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, V) \qquad \text{as } n \to \infty.$$

The theorem follows from a quite general theorem on sequences of random vectors which fulfill certain linear difference equations and whose covariances

v_{ij} are strictly positive. Irreducible aperiodic chains satisfy these requirements. If the chain is periodic, a modified version of Theorem 2.18 can be given. The asymptotic distribution then is composed of several singular ones and a multivariate normal, all of which are independent, see [49, p. 215].

The CLT for Markov chains allows to apply the whole machinery of statistical tests suited to multivariate normal distributions. In the appendix of the aforementioned book, the author considers tests for the randomness of tables of random numbers on the basis of some chain models, for example. We will pick up this idea and present it in a more general form which allows to deduce several well-known test statistics for pseudorandom number generators as special cases in Chapter 3.

Note, that the rank $R(V)$ of the asymptotic covariance matrix V determines the dimension of the limiting multivariate normal distribution of $C^{(n)}$. As mentioned in the last paragraph of Section 2.3.2, this dimension is always less or equal to the dimension of the counter vector $C^{(n)}$ minus one, $R(V) \leq m - 1$.

Corollary 2.19 *Let the chain $(S, \mathbb{P})$, the vector $\mathbf{P}$, and the matrix V be defined as in Theorem 2.18. Further, let $k \in \mathbb{N}$ and $M = (m_{ij})_{(i,j) \in S \times \mathbb{N}_k}$ be a real $m \times k$ matrix. Then*

$$\sqrt{n}\left(\frac{1}{n}C^{(n)}M - \mathbf{P}M\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\emptyset, M'VM\right) \qquad \text{as } n \to \infty.$$

Proof: The class of multivariate normal distributions is closed under linear mappings, see Lemma 6.8 in the Appendix. The corollary thus follows from the Mapping Theorem in R^k. ∎

Corollary 2.19 will prove fruitful in Chapter 3 (Example 3.6) and Chapter 5, where we will make use of linear mappings in order to reduce the dimension of the counter vector $C^{(n)}$.

Chapter 3

Serial Tests and Markov Chains

Serial tests are a backbone in the equidistribution and correlation analysis of pseudorandom number generators (PRNGs). A PRNG produces a sequence $(y_l)_{l\in\mathbb{N}_0}$ of so-called pseudorandom numbers (PRNs) in the unit interval $[0,1)$ which are assumed to behave like a realization of an independent sequence of random variables distributed uniformly on $[0,1)$. We refer the reader to [19] for an introduction.

As an example, consider the linear congruential generator, LCG for short. Defined by integer parameters M, a, b, and $\mathtt{prn}_0$, the LCG$(M,a,b,\mathtt{prn}_0)$ produces a sequence $(\mathtt{prn}_l)_{l\geq 0}$ of integers by $\mathtt{prn}_{l+1} = (a\,\mathtt{prn}_l + b) (\mathrm{mod}\ M)$, i.e. $\mathtt{prn}_{l+1}$ is the integer remainder of dividing $a\,\mathtt{prn}_l + b$ by M. A sequence $(y_l)_{l\geq 0}$ of pseudorandom numbers in $[0,1)$ is defined by $y_l = \mathtt{prn}_l/M$. The LCG's output is a periodic sequence whose period length ϱ is mostly M and depends on the choice of parameters [39, p.169].

Pseudorandom number generators have received increasing interest over the past years since many numerical algorithms depend on a reliable source of pseudo-randomness. For instance, think about the field of stochastic simulation where one is interested in the typical behaviour of a (real-world) system such as a queue in a bank. The state of the system X_n at time n (e.g. the number of customers in the line) is supposed to depend in a deterministic way on a set of parameters Y_i, i in some index set, (Y_i could be the arrival time of the customer i for example) for which no purely deterministic description is available. Based on the assumption that these parameters can

be modelled as random variables with a certain distribution function, the state of the system itself becomes a random variable. If direct mathematical analysis of the properties such as expectation, variance, or distribution function, is not feasible, simulation provides a way of estimating these quantities based on a sample. The PRNs are used as realizations for the parameters Y_i. We will sort of reverse this setup and deduce criteria on the quality of the pseudorandom numbers from known asymptotic properties of a chain model.

Other applications of PRNGs such as numerical integration in higher dimension or optimization seem to involve no randomness at first. Here, randomization can be used to control the ratio between the speed of computation and the accuracy of the result. The value in demand is viewed as expected value of a random variable and thereafter estimated based upon a sample. As for numerical integration, this leads to so-called Monte Carlo methods, as for optimization, we have simulated annealing, genetic algorithms, and others.

In any case, the quality of the chosen source of randomness – the PRNG – can decisively influence the results. Although there exist non-deterministic devices such as amplified noise from a transistor, which play an upcoming role in the field of encryption, the generator is usually implemented by means of a deterministic algorithm which is able to "fake" the required aspects of randomness. Deterministic algorithms have the advantage that their results can be replicated and controlled, and that the modelling and debugging process is simplified significantly. Moreover, they allow rigorous analysis and testing to be performed independently of the actual application. The term *pseudo*-random number generator has been adopted to stress the deterministic provenance of the numbers which are hereafter supposed to behave like realizations of independent random variables distributed uniformly on $[0, 1)$. In the following we will call this assumption H_0. H_0 is in fact no severe restriction since any arbitrary distribution function can be obtained by a transformation of uniform variates. For an implementation and discussion of some recent transformation methods see [51, 52].

Note that due to the finite state space of the computer program which implements the generator, the output of every PRNG becomes periodic, eventually.

In order to be accepted by the scientific community and by the practitioners every PRNG has to undergo an extensive testing procedure. Such tests

are aimed to assure that the generated pseudo randomness cannot be distinguished from "real" randomness in a sense which strongly depends on the target application. As discussed in [18, 29] it is clear that no test can ever prove that a given generator is flawless, because such general purpose generators do not exist: every fixed generator (i.e. every fixed periodic sequence of numbers in the unit interval) will fail in several tests if no further restrictions are imposed on the set of admissible tests, see also Chapters 1 and 2 in [55]. As a consequence, different types of generators are used – for instance – for ciphering clear text and for integrating high-dimensional functions.

However, tests can improve or destroy our confidence into that a generator will do well in *certain* applications. So-called *theoretical tests*, [9, 10, 12, 13, 14, 15, 40] prove properties of the generated sequence of numbers based on a mathematical analysis of the underlying algorithm of the generator. These tests are done a-priori, that is, before any single pseudorandom number has been generated on a computer. They are often used in order to parameterize a generator such like tests for the period length of LCGs given the parameters M, a, and b, for instance. A drawback of theoretical tests is that they usually only allow prognostications of properties of the whole period of PRNs. Since nowadays generators offer period lengths ϱ up to 2^{19937}, such tests tell us only little about the comparatively small samples $y_0, \ldots, y_{n-1}$ that will actually be used in an application of the generator. Regarding the whole period of PRNs as sample space and regarding the seed (i.e. the starting point of the subsequence of length n) as random variable distributed uniformly over the period length, the equidistribution analysis using theoretical tests assures the average quality of samples, however.

Theoretical tests always have to be complemented by *empirical tests* where the sample size n ($n << \varrho$) is chosen with respect to the application, see [2, 5, 28, 30, 32, 46, 53]. Here, the generator is treated as a black box and its behavior with respect to several test statistics is studied. If the distribution of the test statistic is known under H_0, this hypothesis can be tested on a desired level of significance. An empirical test is especially valuable if the used test statistic resembles the target application, see [55, Chapter 2] and the forthcoming [19].

These arguments motivate empirical tests for pseudo-randomness which are based on the simulation of a Markov chain. The chain model permits a kind of flexibility with respect to parameters such as the sample size and the number of possible states of the system – actually, Markov chains are often

used to model certain aspects of a real-world system. The distribution of a suitable test statistic, on the other hand, can be derived from known facts about the long-time behavior of MCs, as we will see below.

In this chapter we recall a standard empirical test known as "Serial test" and model it by counting the number of visits of certain Markov chains in each state in the state space during a given interval of time, see Section 3.1. We complete the construction of a test statistic by applying the chi-square statistic and a generalization thereof in Section 3.2. The chi-square statistic is used to measure the deviation of the counter vector from its expectation. In Chapter 4 we will deal with a more general test statistic for this purpose. For a case study with respect to a well known PRNG we refer the reader to Chapter 5.

3.1 Serial Tests for Pseudorandom Sequences

A Serial test in dimension s consists of three major parts, namely a finite partition of the s-dimensional unit cube, a counter of the number of hits of pseudorandom points in each cell of the partition, and a comparison of the counters to their expectation. In this section we consider the first two parts.

Let $\lambda(\cdot)$ denote the Lebesque measure reduced to $[0,1)$. Choose an integer $m \geq 2$ and let $S = \mathbb{N}_m$. Consider a measurable partition $\mathcal{B} = \{B_1, \ldots, B_m\}$ of the unit interval $[0,1)$ with $\lambda(B_i) > 0$, $i \in S$. For every integer s this induces a partition

$$\mathcal{B}^s = \{B_{i_1} \times \cdots \times B_{i_s} \ : \ (i_1, \ldots, i_s) \in S^s\}$$

of $[0,1)^s$. We call s the dimension of the test. Denote by $(y_l)_{l \geq 0}$ the sequence of pseudorandom numbers as they are produced by the generator. Define the function $g : [0,1) \to \mathbb{N}_m$ by setting $g(y)$ equal to the index i of the set B_i which contains y. Let $(Y_l)_{l \geq 0}$ be a sequence of i.i.d. random variables distributed uniformly on $[0,1)$. This is our model of H_0. We denote the random variables by uppercase letters and use the corresponding lowercase letters for the pseudorandom numbers that substitute the realizations of the random variables.

A Serial test is based on counting the number of pseudorandom points in every set in $\mathcal{B}^s$. The pseudorandom points are represented by s-tuples of pseudorandom numbers from a pseudorandom number generator. We distinguish the following most prominent possibilities to construct these tuples:

the case $s = 1$, the case $s > 1$ with (consecutive) non-overlapping tuples, and the case $s > 1$ with (consecutive) overlapping tuples.

3.1.1 The case $s = 1$

This is an empirical test for the uniformity of the pseudorandom numbers. Let $n \in \mathbb{N}$ be an arbitrary but fixed sample size. Define $Z_l = g(Y_l)$, $l \geq 0$, so that Z_l is an i.i.d. sequence of random variables in S with $P[Z_l = i] = \lambda(B_i)$ for each $i \in S$. Finally let $n \in \mathbb{N}$ and define the vector of counters $C^{(n)} = (C_1^{(n)}, \ldots, C_m^{(n)})$, where $C_i^{(n)} = \#\{l \ : \ Z_l = i, 0 \leq l \leq n-1\}$. Under H_0, this vector is multinomial distributed, $C^{(n)} \sim \mathcal{MN}(n, \mathbf{P})$, with parameters n and $\mathbf{P}$, where

$$\mathbf{P} = (P_1, \ldots P_m) = (\lambda(B_1), \ldots, \lambda(B_m)).$$

As a consequence, the expectation of $C^{(n)}$ is

$$E[C^{(n)}] = n\mathbf{P}, \tag{3.1}$$

and the covariance matrix $V[C^{(n)}]$ is given by

$$V[C^{(n)}] = nV = n(v_{ij})_{(i,j)\in S^2} \tag{3.2}$$

with

$$v_{ij} = \begin{cases} P_i(1-P_i) & \text{for} \quad i = j \\ -P_iP_j & \text{for} \quad i \neq j \end{cases}$$

The normalized vector of counters converges in distribution to a multivariate normal with zero mean and covariance matrix V,

$$\frac{1}{\sqrt{n}}(C^{(n)} - E[C^{(n)}]) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, V).$$

As will be described in Section 3.2, the chi-square statistic can be used to test the hypothesis H_0 for a given sample $(y_0, \ldots, y_{n-1})$ of pseudorandom numbers by rating the deviation of the counter vector $C^{(n)}$ from its expectation. This test checks if the pseudorandom numbers hit each partition about the expected number of times, where the expectation is calculated with respect to the Lebesgue measure. It thus tests the notion of uniformity.

We now reconstruct the test by the aid of a Markov chain. Let the state space of the chain be $S = \{1, \ldots, m\}$, and let $\mathbb{P} = (p_{ij})_{(i,j)\in S^2}$ be defined by

$p_{ij} = P_j = \lambda(B_j)$. Since p_{ij} depends only on the index j this is an example for an independent chain as introduced in Definition 2.2: the information of being in state i at time n does not influence the probability of a visit in state j at time $n+1$. Together with the initial distribution $\mathbf{P}_0$ which we let be equal to $\mathbf{P}$, $(S, \mathbb{P})$ is a model for the sequence (Z_l) which has been defined above since (i) $\mathbf{P}_0$ is a stationary distribution and (ii) the visited states $X_n \in S$ are independent random variables. Here (i) follows from Lemma 2.6 and (ii) follows from Corollary 2.1. Note, that by Lemma 2.4 $(S, \mathbb{P})$ is irreducible and aperiodic.

To finally model the counter vector $C^{(n)}$, we count the number of visits of the chain $(S, \mathbb{P})$ in each state $i \in S$ during the time interval $[0, n-1]$ by letting $C_i^{(n)} = \#\{l \ : \ X_l = i, 0 \leq l \leq n-1\}$ and again put $C^{(n)} = (C_1^{(n)}, \ldots, C_m^{(n)})$.

From the above it is clear that the vector $C^{(n)}$ is multinomial distributed. As an illustrating example for the theory presented in Chapter 2 we nevertheless show the calculation of the asymptotic expectation $\mathbf{E} = \lim_{n\to\infty} \frac{1}{n} E[C^{(n)}]$ and covariance matrix $V = \lim_{n\to\infty} \frac{1}{n} V[C^{(n)}]$ of $C^{(n)}$ by the formulas (2.11), (2.12) and (2.15). The determinants $\mathbb{P}_{ii}(1)$ for every $i \in S$ can be computed by transforming the minor $(I_m - \mathbb{P})_{ii}$ into an upper triangle matrix using column-wise elimination of the elements below the diagonal. The determinant $\mathbb{P}_{ii}(1)$ then equals the product of the diagonal terms in the resulting matrix which gives a telescoping product. Consider at first the case $i = 1$ and recall that $\sum_{i\in S} P_i = 1$, then the according minor calculates to

$$\mathbb{P}_{11}(1) = \frac{P_1}{P_1 + P_2} \cdot \frac{P_1 + P_2}{P_1 + P_2 + P_3} \cdots \frac{P_1 + \ldots + P_{m-2}}{P_1 + \ldots + P_{m-1}} \cdot (1 - P_m) = P_1.$$

By obvious symmetries in $\mathbb{P}$, we immediately have $\mathbb{P}_{ii}(1) = P_i$, so that (2.11) yields $\mathbf{E} = (P_1, \ldots, P_m)$. This is in accordance to the normalized expectation (3.1) of the multinomial distribution and to the fact that $(S, \mathbb{P})$ is independent with respect to $\mathbf{P} = (P_1, \ldots, P_m)$. As to the covariance matrix V, we proceed by calculating the determinant $\mathbb{P}_{ij|ij}(1)$. By a transformation to an upper triangular matrix it is again possible to express this determinant as telescoping product which finally calculates to $\mathbb{P}_{ij|ij}(1) = P_i + P_j$. Owing to $\sum_{k\in S} \mathbb{P}_{kk}(1) = \sum_{k\in S} P_k = 1$ this yields the expressions $Q_{ij} = P_i + P_j$, $Q_i = 1 + (m-2)P_i$, and $Q = 2(m-1)$. By (2.12) and (2.15), $V = (v_{ij})_{(i,j)\in S^2}$ calculates to

$$v_{ij} = \begin{cases} P_i(1 - P_i) & \text{for} \quad i = j \\ -P_i P_j & \text{for} \quad i \neq j \end{cases}$$

Again, this result matches the covariance matrix of a multinomial (3.2). We have thus shown that the additional terms in (2.14) and (2.16) cancel to zero in the case of an independent chain. Since $(S, \mathbb{P})$ is irreducible and aperiodic, Theorem 2.18 gives the convergence in distribution to a multivariate normal,

$$\sqrt{n}\left(\frac{1}{n}C^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, V)\,.$$

We conclude from the above that the Serial test for pseudorandom numbers $(y_0, \ldots, y_{n-1})$ in dimension $s = 1$ amounts to counting the number of visits of an independent chain $(S, \mathbb{P})$ in each state during the time interval $[0, n-1]$. Under H_0 the sequence of partitionated pseudorandom numbers can be viewed as a realization of the process $(X_l)_{l\in\{0,\ldots,n-1\}}$ of states of the chain.

3.1.2 The case $s > 1$ with non-overlapping tuples

The hypothesis H_0 consists of two parts, namely the uniformity of the random variables, and their independence. The latter property is usually tested by measuring the deviation of the empirical distribution of non-overlapping s-tuples $\tilde{y}_l := (y_{ls}, \ldots, y_{ls+s-1})$, $l \geq 0$, from the s-dimensional uniform distribution. To do so, we calculate hits of such points in every set of the partition $\mathcal{B}^s$ of $[0,1)^s$. As in the case $s = 1$, we first define "partitioned" variables $\tilde{Z}_l := (g(Y_{ls}), \ldots, g(Y_{ls+s-1}))$, $\tilde{Z}_l \in S^s$, so that for each $\mathbf{i} = (i_1, \ldots, i_s) \in S^s$,

$$P[\tilde{Z}_l = \mathbf{i}] = \prod_{k=1}^{s} \lambda(B_{i_k}) = \prod_{k=1}^{s} P_{i_k}.$$

An according vector of counters is defined by $\tilde{C}^{(n)} = (\tilde{C}^{(n)}_{\mathbf{i}})_{\mathbf{i}\in S^s}$ with

$$\tilde{C}^{(n)}_{\mathbf{i}} = \#\{l \,:\, \tilde{Z}_l = \mathbf{i}, 0 \leq l \leq n-1\},$$

where we choose the natural with respect to the lexicographical order for the components. Let $\tilde{P} = \Psi(P)$ as defined in Definition 2.8. The hypothesis H_0 and the fact that $\tilde{Z}_l$ is given by *non*-overlapping s-tuples of independent random variables imply the independence of the vectors $\tilde{Z}_l$ so that the vector of counters is again multinomial distributed,

$$\tilde{C}^{(n)} \sim \mathcal{MN}(n, \tilde{P}),$$

with parameters n and $\tilde{P}$.

As in the one-dimensional case, it is easy to calculate the expectation and the covariance matrix of $\tilde{C}^{(n)}$ and to apply a chi-square test to measure the deviation of $\tilde{C}^{(n)}$ from its expectation.

The notation $\tilde{Z}$ already suggested the use of coupled chains for modelling this test by the aid of Markov chains. Let $(S, \mathbb{P})$, $\mathbf{P}$, and $\mathbf{P}_0$ be defined as in the case $s = 1$ and consider the corresponding coupled chain $(\tilde{S}, \tilde{\mathbb{P}})$ with dimension s and initial distribution $\tilde{\mathbf{P}}_0 = \Psi(\mathbf{P}_0)$.

Since $\tilde{\mathbf{P}}_0$ is a stationary distribution by Lemma 2.11 and the coupled chain itself is independent with respect to $\tilde{\mathbf{P}}_0$ by Lemma 2.10, the stochastical equivalence of the sequence $(\tilde{Z}_l)_{l\geq 0}$ and the process $(\tilde{X}_l)_{l\geq 0}$ of states of the chain $(\tilde{S}, \tilde{\mathbb{P}})$ follows readily. By Lemma 2.11 we also get the irreducibility and aperiodicity of $(\tilde{S}, \tilde{\mathbb{P}})$.

Owing to the fact that $(\tilde{S}, \tilde{\mathbb{P}})$ belongs to the class of independent chains, we may use the same calculations as in the case $s = 1$ to express the asymptotic expectation and covariance matrix in terms of minors of the determinant $\tilde{\mathbb{P}}(1)$.

We conclude from the above that the Serial test for pseudorandom numbers $(y_0, \ldots, y_{ns-1})$ in dimension $s > 1$ with non-overlapping tuples amounts to counting the number of visits of a coupled chain with dimension s in each state during the time interval $[0, n-1]$. Under H_0 we can view partitioned consecutive non-overlapping s-tuples of pseudorandom numbers as realization of the process $(\tilde{X}_l)_{l\in\{0,\ldots,n-1\}}$ of states of the coupled chain. The test needs n times s pseudorandom numbers.

3.1.3 The case $s > 1$ with overlapping tuples

It remains to discuss the case where overlapping vectors $\overline{y}_l := (y_l, y_{l+1}, \ldots, y_{l+s-1})$, $l \geq 0$, are considered. The motivation for Serial tests based on such vectors is mainly the same as in the non-overlapping case, that is, a test on the independence presumed by the hypothesis H_0. Numerical practice has shown that the tests tend to reject low-quality generators for about the same sample size n independent of whether overlapping or non-overlapping tuples are used. As we will see below, the case of non-overlapping tuples requires significantly less pseudorandom numbers, which can be a great deal if the generation of the numbers is slow in comparison to the overall speed of the testing procedure.

One has to pay for this benefit with a more sophisticated mathematical model, because the vectors $\overline{Z}_l = (g(Y_l), \ldots, g(Y_{l+s-1}))$ are not independent any more. Define

$$\overline{C}_{\mathbf{i}}^{(n)} = \#\{l \ : \ \overline{Z}_l = \mathbf{i}, 0 \leq l \leq n-1\}$$

and let the counter vector $\overline{C}^{(n)} = (\overline{C}_{\mathbf{i}}^{(n)})_{\mathbf{i} \in S^s}$ where we use the natural with respect to the lexicographical order for the components.

The calculation of the expectation $E[\overline{C}^{(n)}]$ and the covariance matrix $V[\overline{C}^{(n)}]$ are somewhat simplified by considering a cyclic sequence of random variables Y_l, $l \geq 0$, with period length equal to the sample size n, where we let $Y_{l+n} := Y_l$. This changes at most the last $s-1$ vectors $(Y_l, Y_{l+1}, \ldots, Y_{l+s-1})$, $l \in \{n-s+1, n-s+2, \ldots, n-1\}$ and the effect on the standardized counter $\frac{1}{\sqrt{n}}\overline{C}^{(n)}$ thus vanishes for increasing n.

As to the expectation we get for every $\mathbf{i} = (i_1, \ldots, i_s) \in S^s$, $P[\overline{Z}_l = \mathbf{i}] = \prod_{k=1}^{s} P_{i_k}$. Thus

$$\mathbf{E} = \lim_{n \to \infty} \frac{1}{n} E[\overline{C}^{(n)}] = (\prod_{k=1}^{s} P_{i_k})_{\mathbf{i} \in S^s}, \tag{3.3}$$

where we again use the natural with respect to the lexicographical order for the components. As to the covariance matrix $V[\overline{C}^{(n)}]$ it is shown in [16, 55] that $V[\overline{C}^{(n)}] = (\overline{v}_{\mathbf{ij}}^{(n)})$, where $\overline{v}_{\mathbf{ij}}^{(n)} = n\overline{v}_{\mathbf{ij}}$ for a certain matrix $\overline{V} = (\overline{v}_{\mathbf{ij}})$ which depends on $\mathbf{P}$, but not on n. Due to the dependencies arising from the overlapping tuples, the vector of counters cannot be multinomial distributed and we hence cannot apply the chi-square statistic directly. Asymptotic normality,

$$\sqrt{n}(\frac{1}{n}\overline{C}^{(n)} - \mathbf{E}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\emptyset, \overline{V}\right), \tag{3.4}$$

can however be proved by applying e.g. the Central Limit Theorem for s-dependent random variables, see [55, p. 96]. As will be shown in Section 3.2, the difference between two ordinary chi-square statistics, one for dimension s and one for dimension $s-1$, can be used for testing.

We now construct a model for the overlapping Serial test by the aid of an overlapping Markov chain. Let $(S, \mathbb{P})$, $\mathbf{P}$ and $\mathbf{P}_0$ be defined as in the case $s = 1$ and consider the overlapping chain $(\overline{S}, \overline{\mathbb{P}})$ with dimension s and initial distribution $\overline{\mathbf{P}}_0 = \Upsilon(\mathbf{P}_0)$. The stochastical equivalence of the

sequence $(\overline{Z}_l)_{l\geq 0}$ and the process $(\overline{X}_l)_{l\geq 0}$ of states of the chain $(\overline{S}, \overline{\mathbb{P}})$ follows by Lemma 2.12. By Lemma 2.13 we get the irreducibility and aperiodicity of $(\overline{S}, \overline{\mathbb{P}})$.

The matrix calculus for this chain is considerably more complicated. Here we give only a small numerical example.

Example 3.1 *Let* $m = 2$, $S = \{1, 2\}$, $P_1 = P_2 = 1/2$ *and thus* $\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$ *and* $s = 3$, *such that*

$$\overline{\mathbb{P}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Based on a Mathematica implementation of the formulas for $\overline{\mathbb{P}}(1)$ *and its minors, we get the asymptotic expectation*

$$\mathbf{E} = \lim_{n\to\infty} \frac{1}{n} E[\overline{C^{(n)}}] = (\frac{1}{8}, \ldots, \frac{1}{8}),$$

and the asymptotic covariance matrix $V = \lim_{n\to\infty} \frac{1}{n} V[\overline{C^{(n)}}]$, *which calculates to*

$$V = \begin{pmatrix} +\frac{15}{64} & +\frac{1}{64} & -\frac{1}{64} & -\frac{3}{64} & +\frac{1}{64} & -\frac{5}{64} & -\frac{3}{64} & -\frac{5}{64} \\ +\frac{1}{64} & +\frac{3}{64} & +\frac{1}{64} & -\frac{1}{64} & +\frac{3}{64} & -\frac{3}{64} & -\frac{1}{64} & -\frac{3}{64} \\ -\frac{1}{64} & +\frac{1}{64} & +\frac{7}{64} & -\frac{3}{64} & +\frac{1}{64} & +\frac{3}{64} & -\frac{3}{64} & -\frac{5}{64} \\ -\frac{3}{64} & -\frac{1}{64} & -\frac{3}{64} & +\frac{3}{64} & -\frac{1}{64} & +\frac{1}{64} & +\frac{3}{64} & +\frac{1}{64} \\ +\frac{1}{64} & +\frac{3}{64} & +\frac{1}{64} & -\frac{1}{64} & +\frac{3}{64} & -\frac{3}{64} & -\frac{1}{64} & -\frac{3}{64} \\ -\frac{5}{64} & -\frac{3}{64} & +\frac{3}{64} & +\frac{1}{64} & -\frac{3}{64} & +\frac{7}{64} & +\frac{1}{64} & -\frac{1}{64} \\ -\frac{3}{64} & -\frac{1}{64} & -\frac{3}{64} & +\frac{3}{64} & -\frac{1}{64} & +\frac{1}{64} & +\frac{3}{64} & +\frac{1}{64} \\ -\frac{5}{64} & -\frac{3}{64} & -\frac{5}{64} & +\frac{1}{64} & -\frac{3}{64} & -\frac{1}{64} & +\frac{1}{64} & +\frac{15}{64} \end{pmatrix}.$$

As mentioned in Section 2.3.2 this matrix has rank $R(V) = m^s - m^{s-1} = 4$.

We conclude from the above that the Serial test for pseudorandom numbers $(y_0, \ldots, y_{n+s-2})$ in dimension $s > 1$ with overlapping tuples amounts to counting the number of visits of an overlapping chain with dimension s in each state during the time interval $[0, n-1]$. Under H_0 we can view partitioned consecutive overlapping s-tuples of pseudorandom numbers as realization of the process $(\overline{X}_l)_{l \in \{0,\ldots,n-1\}}$ of states of the overlapping chain. The test needs only $n + s - 1$ pseudorandom numbers which is about the s'th fraction in comparison to the Serial test in dimension $s > 1$ with non-overlapping tuples.

We summarize this section in Table 3.1.

dimension	tuples	No. of PRNs	rank $R(V)$
1	-	n	$m-1$
s	non-overlapping	ns	$m^s - 1$
s	overlapping	$n+s-1$	$m^s - m^{s-1}$

Table 3.1: Serial Tests

3.2 Test Statistics for Frequency Analysis

In all the examples of the previous section we have constructed a model based on counting the number of visits of an irreducible aperiodic Markov chain $(S, \mathbb{P})$ in every state $i \in S = \{1, \ldots, m\}$ for a given time interval $[0, n-1]$, where we call n the sample size. Denote this vector of counters by $C^{(n)}$ and denote its expectation and covariance matrix by $E[C^{(n)}]$ and $V[C^{(n)}]$, respectively, and let $\mathbf{E} = \lim_{n\to\infty} \frac{1}{n} E[C^{(n)}]$ and $V = \lim_{n\to\infty} \frac{1}{n} V[C^{(n)}]$ be the asymptotic mean and covariance matrix. Further denote by D_m the set of m-variate discrete probability distributions $D_m = \{\mathbf{P} = (P_1, \ldots, P_m) \in \mathbb{R}^m, P_i \geq 0, i \in \{1, \ldots, m\}, \sum_{i=1}^{m} P_i = 1\}$, and put $D_m^\circ = \{\mathbf{P} \in D_m, P_i > 0, i \in \{1, \ldots, m\}\}$ the set of probability distributions with support $\{1, \ldots, m\}$. From (2.11) and Corollary 2.17 we have $\mathbf{E} = \mathbf{P} \in D_m^\circ$. We let $\hat{\mathbf{P}}^{(n)} = \frac{1}{n} C^{(n)}$ so that $\hat{\mathbf{P}}^{(n)} \in D_m$.

In the case of an independent chain, the vector of counters is multinomial distributed $\mathcal{MN}(n, \mathbf{P})$ with parameters n and $\mathbf{P}$. In the case of an arbitrary

finite irreducible and aperiodic chain, Theorem 2.18 gives the convergence in distribution to a multivariate normal,

$$\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, V) \qquad \text{as } n \to \infty.$$

In this section we complete the setup of Serial tests by applying an adequate goodness-of-fit statistic to rate the deviation of the counter vector from its expectation.

One might tentatively consider the square of the Euclidean norm, $||\hat{\mathbf{P}}^{(n)} - \mathbf{P}||^2$, i.e., the sum of squares of the errors given by the differences between the n'th fraction of the counter vector and its asymptotic expectation.
It is clear on the other hand, that the asymptotic mean $\mathbf{E}$ and covariance matrix V of the counter vector have to be taken into account in order make the asymptotic distribution of the test statistic independent of $\mathbf{E}$ and V and to assure that it belongs to a fixed class of distributions indexed by the dimension of the multivariate normal only: both, a counter with high expectation and a set of two strongly correlated counters, should be weighted less than average for a proper notion of distance.

In this section we thus recall two well known weighting procedures, namely the chi-square statistic of Karl Pearson, and a generalization thereof, so-called quadratic forms in weak inverses.

3.2.1 Pearson's Statistic for Multinomial Variates

A well-known test for the hypothesis $\tilde{H}_0$ that the vector $C^{(n)}$ is distributed multinomial with parameters n and $\mathbf{P} = (P_1, \ldots, P_m)$ is the so-called chi-square statistic $\mathcal{X}^2$ of K. Pearson,

$$\mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = n \sum_{i=1}^{m} \frac{(\hat{P}_i^{(n)} - P_i)^2}{P_i}.$$

Here we have weights $1/P_i$ which are aimed to assure an asymptotic distribution of $\mathcal{X}^2$ which is independent of $\mathbf{P}$. The goal is achieved in fact since under $\tilde{H}_0$, the asymptotic distribution of $\mathcal{X}^2$ as n goes to infinity is a chi-square with $m-1$ degrees of freedom for every $\mathbf{P} \in D_m^\circ$. This is the statement of the famous Theorem of Pearson [45].

Since the counter of a Serial test in dimension $s = 1$ (Section 3.1.1) and in dimension $s > 1$ with non-overlapping tuples (Section 3.1.2) is multinomial

distributed under the hypothesis H_0 that the pseudorandom numbers are sampled from a sequence of independent random variables distributed uniformly on $[0,1)$, we can test H_0 with the statistic $\mathcal{X}^2$ as follows: let χ^2_{m-1} denote a chi-square random variable with $m-1$ degrees of freedom. We fix a sample size n and calculate the vectors $\hat{\mathbf{P}}^{(n)}$ and $\mathbf{P}$. Given a rejection region $\mathcal{C}$ with $P[\chi^2_{m-1} \in \mathcal{C}] = \alpha > 0$, H_0 is rejected at the level of significance α if $\mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \in \mathcal{C}$.

Note, that if the counter vector of a Serial test in dimension $s = 1$ has dimension m (corresponding to the cardinality $m = \#S$ of the state space of the according chain), the counter vector for the Serial test in dimension $s > 1$ with non-overlapping tuples has dimension m^s. This value has to be substituted for m in the calculation of the number of degrees of freedom in the above decision rule. In both cases, the number of degrees of freedom equals the rank $R(V)$ of the asymptotic covariance matrix V of the counter vector, see also Section 2.3.2.

Attention has to be paid to the fact that the distribution of $\mathcal{X}^2$ is a chi-square only in the asymptotic case $n \to \infty$. Rules of thumb for the quality of the approximation with respect to finite n and given $\mathbf{P}$ can be found in [17, 182], for example.

3.2.2 Quadratic Forms in Weak Inverses

Throughout this section we let Σ a $m \times m$ covariance matrix with rank $r = R(\Sigma)$ and assume a vector $\mathbf{P} \in D_m^\circ$ and a sequence $(\hat{\mathbf{P}}^{(n)})_{n\in\mathbb{N}}$, $\hat{\mathbf{P}}^{(n)} \in D_m$, of random vectors which satisfies

$$\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)\,.$$

Let $S = \{1, \ldots, m\}$ and denote by $\overline{\Sigma} = (\overline{\sigma}_{ij})_{(i,j)\in S^2}$ a real $m \times m$ matrix. We define the statistic

$$\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) := n \sum_{i,j=1}^{m} \overline{\sigma}_{ij} \left(\hat{P}_i^{(n)} - P_i\right)\left(\hat{P}_j^{(n)} - P_j\right). \tag{3.5}$$

In matrix notation, $\mathcal{X}^2_{\overline{\Sigma}}$ assumes the form of a quadratic form in the matrix $\overline{\Sigma}$, $\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = n(\hat{\mathbf{P}}^{(n)} - \mathbf{P})\overline{\Sigma}(\hat{\mathbf{P}}^{(n)} - \mathbf{P})'$. Below we give a criterion on $\overline{\Sigma}$ based on the concept of weak inverses of the matrix Σ which assures, that the asymptotic distribution of $\mathcal{X}^2_{\overline{\Sigma}}$ is a chi-square with r degrees of freedom.

The statistic can be used in the case of a Serial test in dimension $s > 1$ with overlapping tuples, for example.

Definition 3.1 *Let A be a real $l \times k$ matrix. If there exists an $k \times l$ matrix A^- with $AA^-A = A$, this matrix is called weak inverse of A.*

In some literature, the term generalized inverse is used. The theory of generalized inverses is a broad field in linear algebra. Such inverses can be constructed in different ways, see [47, 20, 37]. As B^{-1} does in the case of regular matrices B, A^- serves as an "equation solver" for arbitrary, not necessarily regular matrices A: let $y \in \mathbb{R}^k$ be an element in the range $A(\mathbb{R}^l)$ of A such that there exists a $z \in \mathbb{R}^l$ with $Az = y$. Then $x := A^-y$ is a solution of the equation $Ax = y$. The proof is simple and therefore omitted.

The construction of a weak inverse Σ^- of a covariance matrix Σ can be done by applying the spectral decomposition. Let Σ be a real, symmetric $m \times m$ matrix, $\Sigma = \Sigma'$, and denote the rank of Σ by $r = R(\Sigma)$. Here Σ' denotes the transposed matrix. Since Σ is real and symmetric, there exists an orthogonal matrix U such that $U'\Sigma U = D$, where $D = diag(\lambda_1, \ldots, \lambda_m)$ is a diagonal matrix containing the (not necessarily distinct) eigenvalues $\lambda_i \in \mathbb{R}$ of Σ. Since $UU' = I$ it follows that $\Sigma = UDU'$. Now let

$$\lambda_i^- = \begin{cases} 1/\lambda_i & \text{for} \quad \lambda_i \neq 0 \\ 0 & \text{for} \quad \lambda_i = 0 \end{cases},$$

and set $D^- = diag(\lambda_1^-, \ldots, \lambda_m^-)$, and $\Sigma^- := UD^-U'$. Then

$$\Sigma\Sigma^-\Sigma = UDU'UD^-U'UDU' = UDU' = \Sigma,$$

proving that Σ^- is a weak inverse of Σ. This assures the existence of a weak inverse for arbitrary covariance matrices.

To return to our main topic, namely the distribution of the statistic $\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, we cite [47, Theorem 9.2.2] without proof.

Theorem 3.2 (Distribution of quadratic forms) *Let the random vector X be multivariate normal distributed, $X \sim \mathcal{N}(\emptyset, \Sigma)$, with zero mean and covariance matrix Σ. Furthermore, let $\overline{\Sigma} = \Sigma^-$ be a weak inverse of Σ and let $r = R(\Sigma)$ be the rank of Σ. Then the statistic $X\overline{\Sigma}X'$ is chi-square distributed with r degrees of freedom,*

$$X\overline{\Sigma}X' \sim \chi^2_r.$$

The weak inverse Σ^- of the covariance matrix Σ thus turns out to be the proper system of weights $\overline{\Sigma}$ such that the statistic (3.5), which thereby becomes a quadratic form in a weak inverse, has a limiting distribution from the class of chi-square distributions:

Corollary 3.3 *If* $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$ *then* $\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\mathrm{d}} \chi^2_r$.

The Corollary follows by the Mapping Theorem in $\mathbb{R}^m$. ■

We give the following three examples for instances of $\mathcal{X}^2_{\overline{\Sigma}}$: the chi-square statistic of Pearson itself, which thus becomes a special case of (3.5), the overlapping Serial test statistic, and a combination of the overlapping Serial test with a linear mapping that reduces the dimension of the counter vector:

Example 3.4 (The Pearson chi-square) *Let* $S = \{1, \ldots, m\}$ *and let* $n\hat{\mathbf{P}}^{(n)}$ *be multinomial distributed,* $n\hat{\mathbf{P}}^{(n)} \sim \mathcal{MN}(n, \mathbf{P})$, *with parameters* n *and* $\mathbf{P} = (P_1, \ldots, P_m)$, *where* $P_i > 0$ *for* $i \in S$, *so that* $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$ *where* $\Sigma = (\sigma_{ij})_{(i,j)\in S^2}$ *equals the covariance matrix of the multinomial distribution* $\mathcal{MN}(1, \mathbf{P})$,

$$\sigma_{ij} = \begin{cases} P_i(1 - P_i) & \text{for} \quad i = j \\ -P_i P_j & \text{for} \quad i \neq j \end{cases},$$

which has rank $r = R(\Sigma) = m - 1$. *A weak inverse for* Σ *is given by* $\Sigma^- = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$, *as can be proved by straight forward calculation of* $\Sigma\Sigma^-\Sigma = \Sigma$, *see Lemma 6.9 in the Appendix. Substituting* $\overline{\Sigma} = \Sigma^-$ *in (3.5), the non-diagonal terms cancel and we get exactly Pearson's statistic* $\mathcal{X}^2$. *The number of degrees of freedom equals the rank* r *of* Σ. *The convergence of the test statistic to a chi-square distribution thus results from the convergence of the multinomial to a multivariate normal in this case.*

Example 3.5 (The overlapping Serial test) *Recall the case* $s > 1$ *with overlapping tuples from Section 3.1.3: Let* $(S, \mathbb{P})$ *be the independent chain with respect to* $\mathbf{P} = (P_1, \ldots, P_m)$, *where* $P_i > 0$ *for* $i \in S$. *Consider the counter vector* $\overline{C}^{(n)}$ *of the overlapping chain* $(\overline{S}, \overline{\mathbb{P}})$ *with dimension* $s > 1$ *of* $(S, \mathbb{P})$. *Let* $\mathbf{E}$ *be defined as in (3.3). From (3.4) we have asymptotic normality,*

$$\sqrt{n}(\frac{1}{n}\overline{C}^{(n)} - \mathbf{E}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\emptyset, \overline{V}\right),$$

of the counter vector $\overline{C}^{(n)}$. A weak inverse $\overline{V}^-$ for the covariance matrix $\overline{V}$ is given in e.g. [55, p. 108]. Substituting $\overline{\Sigma} = \overline{V}^-$ in (3.5), it turns out that $\mathcal{X}^2_{\overline{\Sigma}}$ equals the difference of two ordinary chi-square statistics, one for dimension s and one for dimension $s-1$. This has already been proved by Good [16] in 1953 by a quite different approach. The rank r of $\overline{V}^-$ equals $r = m^s - m^{s-1}$, see Section 2.3.2.
The overlapping Serial test for pseudorandom numbers $y_0, \ldots, y_{n+s-2}$ in dimension $s > 1$ can thus be computed as follows: let χ^2_r denote a chi-square random variable with r degrees of freedom. Calculate the counter vector $\overline{C}^{(n)}$ based on the pseudorandom numbers as described in Section 3.1.3, let $\hat{\mathbf{P}}^{(n)} = \frac{1}{n}\overline{C}^{(n)}$ and compute the value of $\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{E})$. Given a rejection region $\mathcal{C}$ with significance level $P[\chi^2_r \in \mathcal{C}] = \alpha$, $0 < \alpha \ll 1$, H_0 is rejected at the level of significance α if $\mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{E}) \in \mathcal{C}$. Again, the convergence to a chi-square distribution stems from the convergence of the vector of counters to a multivariate normal. Attention has therefore to be paid to the approximation error for finite n.

Example 3.6 (Dimension Reduction) *A major difficulty in performing the overlapping Serial test is the cardinality of $\overline{S}$ which equals m^s since $\mathbb{P}$ is strictly positive, see Section 5.2 for a more concrete example. As a way out of this problem we consider a linear mapping to reduce the dimension of the counter vector $\overline{C}^{(n)}$. Let $\mathbf{P}$, $(S, \mathbb{P})$, $(\overline{S}, \overline{\mathbb{P}})$, $\mathbf{E}$, and $\overline{V}$ be defined as in Example 3.5. For an integer $k \ll m^s$ put M a non-degenerate (i.e. non zero) real $m^s \times k$ matrix and let $\tilde{C}^{(n)} := \overline{C}^{(n)} M \in \mathbb{R}^k$. Do not confuse this notation with that of the coupled chains in Section 2.2.1. By Corollary 2.19 we get $\sqrt{n}(\frac{1}{n}\tilde{C}^{(n)} - M\mathbf{E}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\emptyset, M'\overline{V}M\right)$, with covariance matrix $\tilde{V} := M'\overline{V}M$. For the application of the test statistic $\mathcal{X}^2_{\overline{\Sigma}}$ it remains to calculate a weak inverse $\tilde{V}^-$ and the rank $R(\tilde{V})$ of $\tilde{V}$, see again Section 5.2 for a more concrete example.*

It is especially convenient to construct M by partitioning $\overline{S}$ into k mutually distinct sets $\mathcal{S}_1, \ldots, \mathcal{S}_k$ and by setting

$$M = (m_{ij})_{(i,j) \in \overline{S} \times \mathbb{N}_k} \quad \text{with} \quad m_{ij} := \mathbf{1}_{\mathcal{S}_j}(i). \tag{3.6}$$

By this, the component $\tilde{C}^{(n)}_j$ of $\tilde{C}^{(n)}$ equals the sum $\sum_{i \in \mathcal{S}_j} \overline{C}^{(n)}_i$ of components of the original counter.

In brief, coupling the asymptotic theory for Markov chains with the theory of the distribution of quadratic forms in weak inverses results in a huge class of test statistics for which we can deduce the asymptotic distribution. Like standard Serial tests, these tests are based on the frequency count in the chain. The asymptotic distribution is a chi-square independent of the transition matrix and initial probabilities of the chain. The number of degrees of freedom equals the rank of the asymptotic covariance matrix which is also the dimension of the asymptotic multivariate normal distribution of the counter vector. For the construction of the test statistic, we need to compute the asymptotic expectation and a weak inverse of the asymptotic covariance matrix of the vector of counters for the number of visits in each state of the chain. We may use a linear mapping to reduce the dimension of the counter vector.

The theory covers a wide range of well-known test statistics used in the assessment of pseudorandom number generators. In particular, we have modelled the class of Serial tests in dimension one and in dimensions greater than one with both, overlapping and non-overlapping tuples.

We have seen that a famous measure suited to samples from a multinomial distribution, the Pearson chi-square, can be generalized to work with samples from arbitrary asymptotic multivariate normal distributions. This corresponds to the transition from independent Markov chains to arbitrary finite state Markov chains. It remains to discuss the applicability of other famous distance measures for the multinomial setting – like the log-likelihood ratio statistic or the Hellinger distance – in the case of asymptotic multivariate normal distributions. We will undertake the necessary modifications in the next chapter.

Chapter 4

A Generalized ϕ-Divergence

In 1963, Csiszár [8] defined a measure for the deviation of two probability densities $\hat{\mathbf{P}}$ and $\mathbf{P}$, which we will assume to be discrete. His so-called φ-divergence was also introduced independently by Ali and Silvey [1] in 1966. Regarding a multinomial test setting, the original measure has to be scaled in order to obtain convergence in distribution to a chi-square distribution. This scaled φ-divergence may be viewed as a generalization of Pearson's statistic $\mathcal{X}^2$.

This chapter deals with the definition of a generalized ϕ-divergence $I_{\overline{\Sigma},\phi}$ which extends the applicability of φ-divergences to non-multinomial test settings. The new measure will include Pearson's statistic, the ordinary φ-divergence, and quadratic forms in weak inverses as special cases. We will derive the asymptotic distribution of $I_{\overline{\Sigma},\phi}$ under the assumption of asymptotic multivariate normality of a standardized sequence of random vectors $\hat{\mathbf{P}}^{(n)}$.

There exist several alternative ways to define test statistics for asymptotically multivariate normal random vectors with known covariance matrix, which may be based on parameter estimation, for instance. In contrast, we concentrate on the definition of a concept which includes the well-known statistics and extends their applicability.

4.1 From Pearson's Statistic to the φ-divergence

Let us recall the definition of Pearson's statistic $\mathcal{X}^2$ as it has already been introduced in Chapter 3. As in Chapter 3, denote by D_m the set of discrete probability distributions $D_m = \{\mathbf{P} = (P_1, \ldots, P_m) \in \mathbb{R}^m, P_i \geq 0, i \in \{1, \ldots, m\}, \sum_{i=1}^{m} P_i = 1\}$, and put $D_m^\circ = \{\mathbf{P} \in D_m, P_i > 0, i \in \{1, \ldots, m\}\}$ the set of non-degenerate probability distributions in D_m. Let $\mathbf{P} \in D_m^\circ$ and $\hat{\mathbf{P}} \in D_m$, respectively. Pearson's statistic $\mathcal{X}^2$,

$$\mathcal{X}^2(\hat{\mathbf{P}}, \mathbf{P}) = n \sum_{i=1}^{m} \frac{(\hat{P}_i - P_i)^2}{P_i} = n \sum_{i=1}^{m} P_i \left(\frac{\hat{P}_i}{P_i} - 1 \right)^2, \tag{4.1}$$

is used to measure the deviation between $\mathbf{P}$ and $\hat{\mathbf{P}}$. It is easy to see that $\mathcal{X}^2(\hat{\mathbf{P}}, \mathbf{P}) \geq 0$ and equals zero iff $\hat{\mathbf{P}} = \mathbf{P}$. However, the triangular inequality is not satisfied in general, so that $\mathcal{X}^2$ defines no distance in the strict sense. The famous Theorem of Pearson [45] gives the asymptotics of $\mathcal{X}^2$ in the multinomial test setting: if $n\hat{\mathbf{P}}^{(n)}$ is distributed m-variate multinomial with parameters n and $\mathbf{P}$, $n\hat{\mathbf{P}}^{(n)} \sim \mathcal{MN}(n, \mathbf{P})$, then $\mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ converges to a chi-square distribution with $m - 1$ degrees of freedom as n approaches infinity.

The middle term in (4.1) suggest that $\mathcal{X}^2$ may be interpreted as a weighted average of the squared deviations $\hat{P}_i - P_i$, where the weights are given by $(P_i)^{-1}$. These weights equal the diagonal elements in a weak inverse $\overline{\Sigma} = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$ of the covariance matrix of the multinomial distribution $\mathcal{MN}(1, \mathbf{P})$. By this, $\mathcal{X}^2$ becomes a special case of the general quadratic form in the weak inverse $\overline{\Sigma}$,

$$\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}, \mathbf{P}) = n \sum_{i,j=1}^{m} \overline{\Sigma}_{ij} \left(\hat{P}_i - P_i \right) \left(\hat{P}_j - P_j \right). \tag{4.2}$$

As we have discussed in Chapter 3, such quadratic forms generalize $\mathcal{X}^2$ to the case of asymptotic multivariate normal models with known covariance matrix Σ, where $\mathcal{X}^2_{\overline{\Sigma}}$ obeys a chi-square distribution with the rank of Σ degrees of freedom in the limit.

We might on the other hand consider the right hand term in (4.1) and interpret $(u-1)^2$ as a function which rates the likelihood-ratio $\hat{P}_i/P_i$. Again, $\mathcal{X}^2$ is a weighted average of these, where the weights now equal P_i. This

suggests a generalization by considering weighted averages of other functions φ of $\hat{P}_i/P_i$ in order to define a measure for the deviation of $\hat{\mathbf{P}}$ and $\mathbf{P}$. In 1963, Csiszár [8] defined the so-called φ-divergence of two probability densities. We will only consider the case of discrete distributions. The φ-divergence of $\hat{\mathbf{P}}$ and $\mathbf{P}$ is defined by

$$I_\varphi^\circ(\hat{\mathbf{P}}, \mathbf{P}) = \sum_{i=1}^{m} P_i \; \varphi\left(\frac{\hat{P}_i}{P_i}\right). \tag{4.3}$$

Here, $\varphi : [0, \infty) \to (-\infty, \infty]$ is a function which is convex on $(0, \infty)$ and continuous in 0. Depending on the application, one will impose different restrictions on φ. Within the theory of asymptotic distributions of I_φ° it is usual to do without the convexity on $(0, \infty)$ but to assume that φ allows second continuous derivative on some interval $I_\delta := (1-\delta, 1+\delta)$ and satisfies $\varphi''(1) > 0$. The latter is the crucial assumption in order to get an asymptotic distribution which is independent of $\mathbf{P}$. With the notion of a distance measure in mind, we further claim that $\varphi(1) = 0$. It is no loss of generality to assume $\varphi'(1) = 0$ in addition, since for every φ

$$\tilde{\varphi}(u) = \varphi(u) - \varphi'(1)(u-1), \tag{4.4}$$

satisfies $\tilde{\varphi}'(1) = 0$ and, owing to $\sum_{i=1}^{m} \hat{P}_i^{(n)}(\frac{\hat{P}_i^{(n)}}{P_i} - 1) = 0$, $I_\varphi^\circ(\hat{\mathbf{P}}, \mathbf{P})$ equals $I_{\tilde{\varphi}}^\circ(\hat{\mathbf{P}}, \mathbf{P})$ for all values of $\hat{\mathbf{P}}^{(n)}$ and $\mathbf{P}$. Note that the function φ is not restricted in any sense outside the interval I_δ.

A class of examples satisfying these requirements is given by

$$\varphi(u) = \sum_{k=0}^{\infty} c_k (u-1)^k, c_k \in \mathbb{R}, c_0 = c_1 = 0, c_2 > 0 \tag{4.5}$$

such that the series converges for each $u \in [0, \infty)$.

In order to study convergence in distribution to a chi-square in the aforementioned multinomial test setting, we have to scale I_φ° and define

$$I_\varphi(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \frac{2n}{\varphi''(1)} \sum_{i=1}^{m} P_i \; \varphi\left(\frac{\hat{P}_i^{(n)}}{P_i}\right). \tag{4.6}$$

The family of φ-divergences includes several well-known measures for the deviation of two probability distributions, some of which are:

- the class φ_α introduced by Liese and Vajda [31],

$$\varphi_\alpha(u) = \begin{cases} u - 1 - \ln u & \alpha = 0 \\ \frac{\alpha u + 1 - \alpha - u^\alpha}{\alpha(1-\alpha)} & \alpha \in \mathbb{R} \setminus \{0, 1\} \\ 1 - u + u \ln u & \alpha = 1 \end{cases}$$

This class is also known under the name power-divergence in [48, 7], where the family is indexed by a parameter λ which equals $\alpha - 1$. For every member of the class $\varphi_\alpha(u)$, $\alpha \in \mathbb{R}$, we have $\varphi(1) = \varphi'(1) = 0$, and $\varphi''(1) = 1$. The following instances of φ_α play a major role in estimation and decision theory. For $\alpha = 2$ we get $\varphi_2(u) = \frac{1}{2}(u-1)^2$ which corresponds to the Pearson $\mathcal{X}^2$. Choosing $\alpha = -1$ on the other hand yields Neyman's [38] modified chi-square statistic $NM^2 = n \sum_{i=1}^{m} \frac{(\hat{P}_i - P_i)^2}{\hat{P}_i}$, which is equivalent to exchanging $\hat{\mathbf{P}}$ and $\mathbf{P}$ in $\mathcal{X}^2$. For $\alpha = 1$ we get the I-Divergence of Kullback-Leibler [26], $G^2 = 2n \sum_{i=1}^{m} \hat{P}_i \ln(\frac{\hat{P}_i}{P_i})$, which is also called log-likelihood ratio statistic. In the equiprobable case $\mathbf{P} = (\frac{1}{m}, \ldots, \frac{1}{m})$, $G^2/2n$ is equal to $(\ln(m) - \ln(2) H(\hat{\mathbf{P}}))$, where $H(\hat{\mathbf{P}}) = -\sum_{i=1}^{m} \hat{P}_i \log_2(\hat{P}_i)$ is the so-called sample entropy. The case $\alpha = 0$ gives the modified log-likelihood ratio statistic considered by Kullback [25, 24], $GM^2 = 2n \sum_{i=1}^{m} P_i \ln(\frac{P_i}{\hat{P}_i})$. Finally, setting $\alpha = 1/2$ yields $\varphi_{1/2}(u) = 2(\sqrt{u} - 1)^2$ and hence the square of the Hellinger-Distance [35], $F^2 = 4n \sum_{i=1}^{m} (\sqrt{\hat{P}_i} - \sqrt{P_i})^2$, see also [27, Chapter 4, p. 46].

- the class h_α defined by Boekee [4],

$$h_\alpha(u) = \begin{cases} |u^\alpha - 1|^{1/\alpha} & \alpha \in (0, 1] \\ |u - 1|^\alpha & \alpha \in (1, \infty) \end{cases}$$

For $\alpha = 2$ and $\alpha = 1/2$, h_α corresponds to a multiple of φ_α. The standard theory for asymptotic distributions of φ-divergences works only in these cases, although every $\alpha \in (0, 1]$ allows to define a distance of probability distributions in terms of the corresponding φ-divergence.

- the class ϕ_κ as defined in [22],

$$\phi_\kappa(u) = \frac{|u - 1|^\kappa}{2(u+1)^{\kappa - 1}}, \ \kappa \in [1, \infty)$$

For each measure in this class, the corresponding φ-divergence allows the definition of a distance of probability distributions. Asymptotic theory, however, is applicable only in the case $\kappa = 2$ which was introduced by Vincze [54] and also investigated in Le Cam's book [27, Chapter 4, p. 47]. For $\kappa = 2$ we get $\phi_\kappa(1) = \phi'_\kappa(1) = 0$ and $\phi''_\kappa(1) = 1/2$.

- the class f_p introduced by Österreicher and Vajda [44]: let $\mathbb{R}_+ = (0, \infty)$ and put

$$f_p(u) = \begin{cases} u\ln(u) - (1+u)\ln(1+u) + (1+u)\ln(2) & p = 1 \\ \frac{1}{1-1/p}\left[(1+u^p)^{1/p} - 2^{(1/p)-1}(1+u)\right] & p \in \mathbb{R}_+ \setminus \{1\} \\ |1-u|/2 & p = \infty \end{cases}$$

Similar to φ_α in the case $\alpha = 1$, f_1 can be represented by Shannon's entropy measure. Even more, this representation is not limited to the equiprobable case since for every $\mathbf{P} \in D_m^\circ$ we have $I^\circ_{f_1}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \ln(2)(2H(\frac{\hat{\mathbf{P}}^{(n)}+\mathbf{P}}{2}) - [H(\hat{\mathbf{P}}^{(n)}) + H(\mathbf{P})])$. In the case $p = 1/2$, f_p yields the Hellinger divergence. The case $p = 2$ has a peculiar appeal from the geometric point of view, see [43]. For every $p \in (0, \infty]$, the corresponding φ-divergence allows to define a distance. The asymptotic theory works for $p \in (0, \infty)$, where $f_p(1) = f'_p(1) = 0$ and $f''_p(1) = p2^{1/p-2}$.

In the multinomial setting $n\hat{\mathbf{P}}^{(n)} \sim \mathcal{MN}(n, \mathbf{P})$, $\mathbf{P} \in D_m^\circ$, all the mentioned measures for which the asymptotic theory is applicable are stochastically equivalent in the limit. They all converge to a chi-square distribution with $m-1$ degrees of freedom. This can be shown by representing φ as a Taylor series and thereby achieving a reduction to the Pearson case. However, the generalization so far is possible in the multinomial case only.

Here, the following questions arise naturally. If I_φ is a generalization of $\mathcal{X}^2$ in the multinomial case, is it possible to generalize arbitrary quadratic forms in weak inverses $\mathcal{X}^2_{\overline{\Sigma}}$, i.e. the multivariate normal case, too? Can we define a generalized ϕ-divergence $I_{\overline{\Sigma},\phi}$ which is depending on a function ϕ and includes quadratic forms (4.2) as well as φ-divergences (4.6) as special cases? What is the asymptotic distribution of $I_{\overline{\Sigma},\phi}$? We give answers to these questions in the following sections.

4.2 Generalizing the Quadratic Form

In an attempt to unify (4.2) and (4.6) in a single statistic, we define

$$I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) := n \sum_{i,j=1}^{m} c_{ij}\ P_iP_j\ \phi\left(\frac{\hat{P}_i^{(n)}}{P_i},\frac{\hat{P}_j^{(n)}}{P_j}\right). \tag{4.7}$$

Here, $c_{ij} = c_{ij}(\mathbf{P},\overline{\Sigma},\phi) \in \mathbb{R}$ denote weights which depend on $\mathbf{P}$, $\overline{\Sigma}$, and ϕ. $\phi : [0,\infty)^2 \to [\infty,\infty]$ is a function which takes over the role played by φ in (4.6). Let us assume that ϕ has continuous partial derivatives up to order 2 on an open square I_δ^2 with $I_\delta = (1-\delta, 1+\delta)$ and $0 < \delta < 1$, and denote those by

$$\phi_x(a,b) = \frac{\partial\phi(x,y)}{\partial x}_{|_{(x,y)=(a,b)}}, \quad \phi_y(a,b) = \frac{\partial\phi(x,y)}{\partial y}_{|_{(x,y)=(a,b)}},$$
$$\phi_{xy}(a,b) = \frac{\partial^2\phi(x,y)}{\partial x\partial y}_{|_{(x,y)=(a,b)}}, \quad \phi_{xx}(a,b) = \frac{\partial^2\phi(x,y)}{\partial x^2}_{|_{(x,y)=(a,b)}}, \quad \text{and}$$
$$\phi_{yy}(a,b) = \frac{\partial^2\phi(x,y)}{\partial y^2}_{|_{(x,y)=(a,b)}}.$$

We abbreviate $\phi_x(1,1)$ by ϕ_x. The same notation is used for the other derivatives up to order 2. In analogy to the conditions on φ in I_φ we further assume that $\phi(1,1) = \phi_x = \phi_y = 0$. Note that no restrictions are imposed on ϕ outside the open square I_δ^2.

To derive the asymptotic distribution of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P})$ as $n \to \infty$, we represent ϕ for $(x,y) \in I_\delta^2$ by a Taylor series:

$$\phi(x,y) \quad = \quad \phi(1,1) + \phi_x(x-1) + \phi_y(y-1) +$$
$$+\frac{1}{2}\left\{\phi_{xx}(a,b)(x-1)^2 + 2\phi_{xy}(a,b)(x-1)(y-1) + \phi_{yy}(a,b)(y-1)^2\right\},$$

with $a = 1+\Delta(x-1)$ and $b = 1+\Delta(y-1)$ for suitable $\Delta \in (0,1)$. Let $K_\delta(\mathbf{P}) = \{\hat{\mathbf{P}} \in D_m\ :\ (\frac{\hat{P}_1}{P_1},\ldots,\frac{\hat{P}_m}{P_m}) \in I_\delta^m\}$ and abbreviate $\frac{\hat{P}_i^{(n)}}{P_i} - 1$ by $Q_i^{(n)}$. If $\hat{\mathbf{P}}^{(n)} \in K_\delta(\mathbf{P})$, the Taylor representation can be used for each $\phi(\frac{\hat{P}_i^{(n)}}{P_i},\frac{\hat{P}_j^{(n)}}{P_j})$, $(i,j) \in \{1,\ldots,m\}^2$ and we let

$$a_{ij}^{(n)} = 1 + \Delta_{ij}^{(n)}Q_i^{(n)}, \text{ and } b_{ij}^{(n)} = 1 + \Delta_{ij}^{(n)}Q_j^{(n)} \tag{4.8}$$

for appropriate $\Delta_{ij}^{(n)} \in (0,1)$ and put

$$\begin{aligned}
\epsilon_{xx}^{(n)}(i,j) &= \phi_{xx}(a_{ij},b_{ij}) - \phi_{xx}, \\
\epsilon_{xy}^{(n)}(i,j) &= \phi_{xy}(a_{ij},b_{ij}) - \phi_{xy}, \text{ and} \\
\epsilon_{yy}^{(n)}(i,j) &= \phi_{yy}(a_{ij},b_{ij}) - \phi_{yy}.
\end{aligned}$$

If $\hat{\mathbf{P}}^{(n)} \notin K_\delta(\mathbf{P})$, let $\epsilon_{xx}^{(n)}(i,j) = \epsilon_{xy}^{(n)}(i,j) = \epsilon_{yy}^{(n)}(i,j) = 0$. We will omit the upper indices (n) of $Q_i^{(n)}$, $a_{ij}^{(n)}$, and $b_{ij}^{(n)}$ in the sequel.

Now let T be the Taylor expansion of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P})$ up to the second order terms,

$$T(\hat{\mathbf{P}}^{(n)},\mathbf{P}) = \frac{n}{2}\sum_{i,j=1}^{m} c_{ij}\ P_iP_j\left\{\phi_{xx}Q_i^2 + 2\phi_{xy}Q_iQ_j + \phi_{yy}Q_j^2\right\},$$

so that for $\hat{\mathbf{P}}^{(n)} \in K_\delta(\mathbf{P})$ we can represent (4.7) by

$$\begin{aligned}
I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) &= \frac{n}{2}\sum_{i,j=1}^{m} c_{ij}\ P_iP_j\left\{\phi_{xx}(a_{ij},b_{ij})Q_i^2 + 2\phi_{xy}(a_{ij},b_{ij})Q_iQ_j + \right. \\
&\qquad \left. + \phi_{yy}(a_{ij},b_{ij})Q_j^2\right\} = \\
&= T(\hat{\mathbf{P}}^{(n)},\mathbf{P}) + R(\hat{\mathbf{P}}^{(n)},\mathbf{P}),
\end{aligned}$$

where the remainder term R which vanishes outside $K_\delta(\mathbf{P})$ is given by

$$R(\hat{\mathbf{P}}^{(n)},\mathbf{P}) = \frac{n}{2}\sum_{i,j=1}^{m} c_{ij}\ P_iP_j\ R_{ij}(\hat{\mathbf{P}}^{(n)},\mathbf{P}),$$

where $R_{ij}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) = \epsilon_{xx}^{(n)}(i,j)Q_i^2 + 2\epsilon_{xy}^{(n)}(i,j)Q_iQ_j + \epsilon_{yy}^{(n)}(i,j)Q_j^2$. Using a remaining remainder term U, which vanishes on $K_\delta(\mathbf{P})$ and is simply defined by

$$U(\hat{\mathbf{P}}^{(n)},\mathbf{P}) = \begin{cases} 0 & \text{for} \quad \hat{\mathbf{P}}^{(n)} \in K_\delta(\mathbf{P}) \\ I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) - T(\hat{\mathbf{P}}^{(n)},\mathbf{P}) & \text{for} \quad \hat{\mathbf{P}}^{(n)} \notin K_\delta(\mathbf{P}) \end{cases},$$

(4.7) can finally be written

$$I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)},\mathbf{P}) = T(\hat{\mathbf{P}}^{(n)},\mathbf{P}) + R(\hat{\mathbf{P}}^{(n)},\mathbf{P}) + U(\hat{\mathbf{P}}^{(n)},\mathbf{P}). \qquad (4.9)$$

In Section 6.2 in the Appendix we show that under the assumption $\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \stackrel{\mathrm{d}}{\to} \mathcal{N}(\emptyset, \Sigma)$ both remainder terms $U(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ and $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ converge in probability to zero so that the asymptotic distribution of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ is equal to that of $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$. If furthermore $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ equals the quadratic form $\mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ in a weak inverse of Σ we actually know this asymptotic distribution which is a chi-square with the rank of Σ degrees of freedom. We thus will choose the values c_{ij} in such a way that $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$. Hence let $\Delta = \phi_{xy} + \frac{\phi_{xx}+\phi_{yy}}{2}$, and

$$T(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = n \sum_{i,j=1}^{m} d_{ij} \left(\hat{P}_i^{(n)} - P_i\right)\left(\hat{P}_j^{(n)} - P_j\right), \text{ where}$$

$$d_{ij} = \begin{cases} c_{ii}\Delta + \frac{1}{2P_i} \sum\limits_{k \neq i} P_k(\phi_{xx} c_{ik} + \phi_{yy} c_{ki}) & \text{for} \quad i = j \\ \phi_{xy} c_{ij} & \text{for} \quad i \neq j \end{cases}$$

To get the aforementioned equivalence we now solve $d_{ij} = \overline{\Sigma}_{ij}$ for all $(i,j) \in \{1, \ldots, m\}^2$ under the assumption that $\phi_{xy} \neq 0$ and that $\Delta \neq 0$. This, finally, yields

$$c_{ij}(\mathbf{P}, \Sigma, \phi) = \begin{cases} \frac{1}{\Delta}\left(\overline{\Sigma}_{ii} - \frac{1}{2P_i} \sum\limits_{k \neq i} P_k \left(\frac{\phi_{xx}}{\phi_{xy}}\overline{\Sigma}_{ik} + \frac{\phi_{yy}}{\phi_{xy}}\overline{\Sigma}_{ki}\right)\right) & \text{for} \quad i = j \\ \frac{\overline{\Sigma}_{ij}}{\phi_{xy}} & \text{for} \quad i \neq j \end{cases}$$

The conditions $\phi_{xy} \neq 0$ and $\phi_{xy} + \frac{\phi_{xx}+\phi_{yy}}{2} \neq 0$ are necessary in order to solve the equations for the coefficients c_{ij} in such a way that the quadratic form in the weak inverse $\overline{\Sigma}$ can be implemented. If either of the condition is not satisfied, we get for some $\tilde{c}_i \in \mathbb{R}$

$$T(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \tilde{c}_0 \mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) + n \sum_{i=1}^{m} \tilde{c}_i \left(\hat{P}_i^{(n)} - P_i\right)^2.$$

Here, $\tilde{c}_0 = 0$ if $\phi_{xy} = 0$, and $\tilde{c}_0 = 1$ if $\phi_{xy} \neq 0$. The asymptotic distribution of $I_{\overline{\Sigma},\phi}$ ought not to be a central chi-square in this case which we will therefore not consider any further here.

We summarize our results in the following Theorem.

Theorem 4.1 (The Generalized ϕ-Divergence)
Assume, that the function $\phi : [0,\infty)^2 \to (-\infty, \infty]$ has continuous partial

derivatives up to order 2 on an open square $I_\delta^2 \subset \mathbb{R}^2$ *containing the point* $(1,1)$ *and that* $\phi(1,1) = \phi_x = \phi_y = 0$, $\phi_{xy} \neq 0$, *and* $\Delta = \phi_{xy} + \frac{\phi_{xx}+\phi_{yy}}{2} \neq 0$. *For a fixed* $\mathbf{P} \in D_m^\circ$ *assume a sequence of random vectors* $\hat{\mathbf{P}}^{(n)} \in D_m$, $n \in \mathbb{N}$, *which satisfies* $\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$, *with mean vector* $\emptyset$ *and covariance matrix* Σ *and let* $R(\Sigma)$ *denote the rank of* Σ. *Finally, define the generalized* ϕ*-divergence*

$$I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = n \sum_{i,j=1}^{m} c_{ij}\ P_i P_j\ \phi\left(\frac{\hat{P}_i^{(n)}}{P_i}, \frac{\hat{P}_j^{(n)}}{P_j}\right).$$

If

$$c_{ij}(\mathbf{P}, \Sigma, \phi) = \begin{cases} i = j & : \ \frac{1}{\Delta}\left(\overline{\Sigma}_{i,i} - \frac{1}{2P_i}\sum_{k\neq i} P_k\left(\frac{\phi_{xx}}{\phi_{xy}}\overline{\Sigma}_{i,k} + \frac{\phi_{yy}}{\phi_{xy}}\overline{\Sigma}_{k,i}\right)\right) \\ i \neq j & : \ \frac{\overline{\Sigma}_{ij}}{\phi_{xy}} \end{cases},$$

then $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ *converges in distribution to a chi-square distribution* $\chi^2_{R(\Sigma)}$ *with* $R(\Sigma)$ *degrees of freedom.*

4.3 The $\tilde{I}_{\overline{\Sigma},\varphi}$-Divergence

The original φ-divergence depends on a function φ which is defined on $[0,\infty)$ only. In our definition of $I_{\overline{\Sigma},\phi}$, we have to provide a function ϕ which is defined on $[0,\infty)^2$. As we have seen from the previous section, a basic requirement on ϕ is that the mixed second derivative does not vanish at the point $(1,1)$. The calculation of the constants c_{ij} becomes significantly easier if the derivatives ϕ_{xx} and ϕ_{yy} DO vanish at $(1,1)$, however. This can be established by setting

$$\phi^\varphi(x,y) := 2\varphi\left(\frac{x+y}{2}\right) - \frac{\varphi(x)+\varphi(y)}{2}, \tag{4.10}$$

where $\varphi : [0,\infty) \to (-\infty,\infty]$ is a function with continuous second derivative on some interval $I_\delta = (1-\delta, 1+\delta) \subset [0,\infty)$, for which $\varphi(1) = \varphi'(1) = 0$ and $\varphi'' := \varphi''(1) \neq 0$, and let φ be arbitrary outside of I_δ. Note that these are exactly the same conditions on φ as we have assumed in connection with the definition of the ordinary φ-divergence, (4.6). The conditions on ϕ in Theorem 4.1 are satisfied by ϕ^φ since $\phi^\varphi(1,1) = \varphi(1) = 0$, $\phi_x^\varphi = \phi_y^\varphi =$

$\varphi'(1) = 0$, $\phi^{\varphi}_{xx} = \phi^{\varphi}_{yy} = 0$ and $\phi^{\varphi}_{xy} = \frac{\varphi''}{2} \neq 0$. Trivially, ϕ^{φ} has continuous derivatives of order 2 on an open square $I^2_\delta \subset \mathbb{R}^2$ containing the point $(1,1)$. The constants c_{ij} now calculate to $c_{ij} = \frac{2}{\varphi''}\overline{\Sigma}_{ij}$ for all $(i,j) \in \{1,\ldots,m\}^2$ and we get the following theorem:

Theorem 4.2 (The $\tilde{I}_{\overline{\Sigma},\varphi}$-Divergence) *Let $\varphi : [0,\infty) \to (-\infty,\infty]$ be a function with continuous second derivative on some interval $I_\delta = (1-\delta, 1+\delta) \subset [0,\infty)$, for which $\varphi(1) = \varphi'(1) = 0$ and $\varphi'' := \varphi''(1) \neq 0$, and let φ be arbitrary outside of I_δ. Let ϕ^{φ} be defined as in (4.10) and set*

$$\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \frac{2n}{\varphi''} \sum_{i,j=1}^{m} \overline{\Sigma}_{ij} P_i P_j \phi^{\varphi}\left(\frac{\hat{P}_i^{(n)}}{P_i}, \frac{\hat{P}_j^{(n)}}{P_j}\right).$$

On the condition that $\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$, this statistic is asymptotically distributed chi-square with $R(\Sigma)$ degrees of freedom.

The resulting statistic $\tilde{I}_{\overline{\Sigma},\varphi}$ is much easier to handle and provides full backward compatibility to $\mathcal{X}^2_{\overline{\Sigma}}$, I_φ, and $\mathcal{X}^2$ as we will see below.

The function $\phi^{\varphi}(x,y)$ has been chosen such that mixed derivatives of order 2 vanish at $(1,1)$. Another class of functions with similar properties can be defined in analogy to (4.5). For this purpose let $\phi(x,y) = c(x-1)(y-1) + \tilde{\phi}(x,y)$, where $c \in \mathbb{R} \setminus \{0\}$ is a constant, and $\tilde{\phi}(x,y)$ is a convergent series of the form $\sum_{k,l\in\mathbb{N}} c_{kl}(x-1)^k(y-1)^l$ where $c_{kl} \in \mathbb{R}$, $c_{kl} = 0$ if $k+l \leq 2$. In this case, the normalizing constant $\frac{2n}{\varphi''}$ has to be replaced by $\frac{n}{c}$.

4.4 Backward Compatibility

As stated in the next Lemma, the statistic $\tilde{I}_{\overline{\Sigma},\varphi}$ is the desired generalization of I_φ and $\mathcal{X}^2_{\overline{\Sigma}}$ in the sense mentioned in the introduction of this chapter.

Lemma 4.3 *Let φ, ϕ^{φ}, and $\tilde{I}_{\overline{\Sigma},\varphi}$ be defined as in Theorem 4.2. Then*

i) choosing $\varphi(u) = (u-1)^2$ we get $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, i.e. the new statistic generalizes the quadratic form in a weak inverse,

ii) choosing $\overline{\Sigma} = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$*, which is a weak inverse of the covariance matrix of the multinomial distribution* $\mathcal{MN}(1, \mathbf{P})$ *(see Lemma 6.9 in the Appendix), yields* $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = I_\varphi(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$*, i.e. the new statistic generalizes the* φ*-divergence* I_φ*. Finally*

iii) choosing both $\varphi(u) = (u-1)^2$*, and* $\overline{\Sigma} = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$ *yields* $\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \mathcal{X}^2(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$*, i.e. backward compatibility to Pearson's statistic.*

Proof: As for i), let $\varphi(u) = (u-1)^2$, then $\phi^\varphi(x, y) = (x-1)(y-1)$ and $\varphi'' = 2$ which yields

$$\tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = n \sum_{i,j=1}^{m} \overline{\Sigma}_{ij}(\hat{P}_i^{(n)} - P_i)(\hat{P}_j^{(n)} - P_j) = \mathcal{X}^2_{\overline{\Sigma}}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$$

As for ii), the non-diagonal terms in $\tilde{I}_{\overline{\Sigma},\varphi}$ vanish and we further have $\phi^\varphi(x, x) = \varphi(x)$. Thus for $\overline{\Sigma} = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$ and arbitrary φ

$$\begin{aligned} \tilde{I}_{\overline{\Sigma},\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) &= \frac{2n}{\varphi''} \sum_{i=1}^{m} \frac{1}{P_i} P_i P_i \phi^\varphi \left(\frac{\hat{P}_i^{(n)}}{P_i}, \frac{\hat{P}_i^{(n)}}{P_i} \right) \\ &= \frac{2n}{\varphi''} \sum_{i=1}^{m} P_i \varphi \left(\frac{\hat{P}_i^{(n)}}{P_i} \right) = I_\varphi(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \end{aligned}$$

Trivially, iii) follows from i) or ii). ■

Chapter 5

Examples

Complementing the theory in Chapters 2,3, and 4, we study two examples in which we join together the concepts of modelling with a Markov chain and testing with the generalized ϕ-divergence. Consequently, the two main steps are the construction of the chain and the choice of a divergence measure from the class of $\tilde{I}_{\overline{\Sigma},\varphi}$-divergences. We also show how to apply the dimension reduction technique introduced in Chapter 3 (Example 3.6). It will become clear, that dimension reduction can either be part of the modelling process itself (see Section 5.1) or arise as a necessity during the implementation stage (see Section 5.2).

As an explanatory example we first recall the Bohr Hydrogen Atom in Section 5.1. In Section 5.2 we discuss a more serious example where we introduce a very promising variant of the overlapping-serial-type of empirical tests for pseudorandom number generators. The test will be used to empirically reveal a wide difference between two long-period linear generators, one of which has a (known) defect with respect to the correlation structure in higher dimensions. The standard overlapping Serial test design does not apply here due to computational difficulties.

5.1 Modelling and Testing Bohr's Hydrogen Atom

Bohr's model of the hydrogen atom – as has already been discussed in Chapter 2 – consists of a proton as the atomic nucleus and a single electron. The state of this system at any time is characterized by the orbit of the electron.

We simplify the model in that we assume that transitions among the different orbits may occur only at time points $n \in \mathbb{N}$ on a discrete time scale and construct a chain model for this situation. Let us index the possible orbits with the set $\mathbb{N} \cup \{\infty\}$. The orbit ∞ corresponds to the ionized state where the electron is not bound to the atom. In order to further simplify the model so that we can apply the theory developed in the preceding chapters, we reduce the infinite number of possible states to a finite state space $S = \mathbb{N}_m$, $m \in \mathbb{N}$ "large enough". Let the state k, $k \in \{1, \ldots, m-1\}$ correspond to the orbit with this index number and identify the state m with the set of all orbits with indices in the set $\{m, m+1, \ldots\} \cup \{\infty\}$. With respect to the physical background this simplification means that we do not distinguish orbits where the state of the electron can, according to Bohr's correspondence principle, be described in classical approximation as a harmonic oscillator.

The next step in the construction of a chain model is crucial. We define the laws of change by assuming a matrix $\mathbb{P}$ of transition probabilities between the different states in S. The matrix $\mathbb{P}$ could, for example, be deduced within the framework of quantum mechanics from the dipole approximation for the time-independent transition probabilities between the corresponding states (which differ from the states $i \in S$ in our model). In contrast to Bohr's theory introducing the quantization of certain physical variables such as energy and angular momentum without providing quantitative estimates for the laws of change, quantum mechanics implies – in principle – the quantitative relation between the experimental setup, the time scale, and the transition probabilities $\mathbb{P}$. Bear in mind, that this is only an illustratory example! We do not suggest using the discrete Markov model as a realistic description of the physical processes in the hydrogen atom.

We assume that the resulting chain $(S, \mathbb{P})$ is irreducible and aperiodic here and in the following and denote the sequence of states by $(X_n)_{n \in \mathbb{N}_0}$. Our aim in this example is to test the model $(S, \mathbb{P})$ by means of experiment and statistical reasoning. As for the experiment, let us assume that we are equipped with the necessary tools to "observe" a real-world atom and to "count" the number of occurrences of each possible state during a time interval $[0, n-1]$. The test will be be based on comparing the relative frequencies to the according expectations in our model $(S, \mathbb{P})$.

Recall from Section 4.3, Theorem 4.2, that the fundamental prerequisite for an application of an $\tilde{I}_{\overline{\Sigma},\varphi}$-divergence is an asymptotically multivariate normal distributed sequence of random vectors with known asymptotic expectation $\mathbf{P}$ and covariance matrix V. By the Central Limit Theorem for irreducible

aperiodic Markov chains 2.18, we have exactly such a sequence by considering the random vectors $\sqrt{n}(\frac{1}{n}C^{(n)} - \mathbf{P})$, where $C^{(n)} = (C_1^{(n)}, \ldots, C_m^{(n)})$, with $C_i^{(n)} := \#\{l : X_l = i, 0 \leq l \leq n-1\}$. So the next steps towards the desired test statistic are the calculation of $\mathbf{P}$ and V on the basis of the formulas (2.11), (2.12) and (2.15), and the computation of both, the rank $R(V)$, and a weak inverse V^- of V. This can, for example, be done by *Mathematica* using the built-in functions `Det` and `PseudoInverse`. As a result we have for $\hat{\mathbf{P}}^{(n)} := \frac{1}{n}C^{(n)}$ that $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, V)$ with $VV^-V = V$.

To fully parameterize the $\tilde{I}_{\overline{\Sigma},\varphi}$-divergence we finally need to supply a function φ according to the criteria stated in Section 4.3, that is, $\varphi : [0,\infty) \to (-\infty,\infty]$ with continuous second derivative on some interval $I_\delta = (1-\delta, 1+\delta) \subset [0,\infty)$, and with $\varphi(1) = \varphi'(1) = 0$, $\varphi'' := \varphi''(1) \neq 0$. As an example we consider the log-likelihood ratio from the class φ_α mentioned in Section 4.1, $\varphi(u) = 1 - u + u \ln u$. Following (4.10), the function ϕ^φ becomes

$$\begin{aligned} \phi^\varphi(x,y) &= 2\varphi\left(\frac{x+y}{2}\right) - \frac{\varphi(x)+\varphi(y)}{2} = \\ &= 1 - \frac{x+y}{2} + x\ln\left(\frac{x+y}{2\sqrt{x}}\right) + y\ln\left(\frac{x+y}{2\sqrt{y}}\right), \end{aligned}$$

in terms of which we get the generalized log-likelihood divergence

$$\tilde{I}_{V^-,\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = 2n \sum_{(i,j)\in S^2} V_{ij}^- P_i P_j \phi^\varphi\left(\frac{\hat{P}_i^{(n)}}{P_i}, \frac{\hat{P}_j^{(n)}}{P_j}\right).$$

This statistic is asymptotically chi-square distributed with $r = R(V)$ degrees of freedom as $n \to \infty$. Provided that n is large enough so that the approximation error is negligible, we can empirically assess our assumptions on the model $(S, \mathbb{P})$. Let χ_r^2 denote a chi-square distributed random variable with r degrees of freedom. Given a rejection region $\mathcal{C}$ with $P[\chi_r^2 \in \mathcal{C}] = \alpha > 0$, we reject the null-hypothesis that $(S, \mathbb{P})$ is a model for the observed states of the hydrogen atom at the level of significance α if $\tilde{I}_{V^-,\varphi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \in \mathcal{C}$.

Note, that the initial distribution $\mathbf{P}_0$ for the chain model $(S, \mathbb{P})$ has asymptotically negligible influence on the test results since the chain forgets this information at an exponential rate due to the finite state space and the irreducibility. $\mathbf{P}_0$ does not appear in any of the formulas (2.11) and (2.12)–(2.16) consistently.

5.1.1 From States to Transitions

In the above example we need information about the states of the electron of a single hydrogen atom during a certain interval of time. It is indeed possible to observe these different states $X_n \in S$, $n \in \mathbb{N}_0$, directly by photoelectron spectroscopy, whereas by traditional optical spectroscopy only energy differences between a state i and a state $j \neq i$ can be detected. Such a transition is accompanied by the absorption or emission of a photon with a specific energy. We denote the transition by the pair (i, j) of successive states. With an optical spectrometer we can in principle measure the intensity of each such energy, that is, we can measure the occurrence rate of the corresponding transition by averaging over a certain number of atoms and/or over some time interval $[0, n-1]$. This corresponds to the experimental determination of the relative frequencies of transitions of the form $(i, j) \in S^2$, $j \neq i$. From these and from known facts about the stability of the states (i.e. the average time an electron spends in a certain state until changing to a state with lower or higher energy level), we can estimate the average occurrence rate of the "transitions" of the form (i, i), $i \in S$, in our model. So, for the following test setup, assume that we have obtained a set of relative frequencies of the number of occurrences of the transitions $\{(i, j) \in S^2\}$.

A model for these frequencies is given by the n'th fraction of the counter vector $\overline{C}^{(n)}$ defined on the overlapping chain $(\overline{S}, \overline{\mathbb{P}})$ with dimension 2 of $(S, \mathbb{P})$. Since $(S, \mathbb{P})$ is irreducible and aperiodic it follows from Lemma 2.13 that $(\overline{S}, \overline{\mathbb{P}})$ is irreducible and aperiodic, too. Thus the Central Limit Theorem applies again and we can proceed like in the above case to construct a generalized log-likelihood divergence for the counter $\overline{C}^{(n)}$. To be more specific, we denote the asymptotic expectation $\mathbf{E} = \lim_{n\to\infty} \frac{1}{n} E[\overline{C}^{(n)}]$ of $\overline{C}^{(n)}$ by $\overline{\mathbf{P}}$, the asymptotic covariance matrix $\lim_{n\to\infty} \frac{1}{n} V[\overline{C}^{(n)}]$ by $\overline{V}$ and a weak inverse of $\overline{V}$ by $\overline{V}^-$. By Theorem 2.18 we have $\sqrt{n}(\frac{1}{n}\overline{C}^{(n)} - \overline{\mathbf{P}}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\emptyset, \overline{V}\right)$ and the generalized log-likelihood divergence $\tilde{I}_{\overline{V}^-,\varphi}(\frac{1}{n}\overline{C}^{(n)}, \overline{\mathbf{P}})$ converges to a chi-square distribution with $R(\overline{V})$ degrees of freedom,

$$\tilde{I}_{\overline{V}^-,\varphi}(\frac{1}{n}\overline{C}^{(n)}, \overline{\mathbf{P}}) = 2n \sum_{(\mathbf{i},\mathbf{j})\in\overline{S}^2} \overline{V}^-_{\mathbf{ij}} \mathbf{P_i} \mathbf{P_j} \phi^{\varphi}\left(\frac{\frac{1}{n}\overline{C}^{(n)}_{\mathbf{i}}}{\overline{\mathbf{P}}_{\mathbf{i}}}, \frac{\frac{1}{n}\overline{C}^{(n)}_{\mathbf{j}}}{\overline{\mathbf{P}}_{\mathbf{j}}}\right) \xrightarrow{\mathrm{d}} \chi^2_{R(\overline{V})}.$$

By using the overlapping chain we have overcome one main drawback of the

original test design for the assessment of the chain model for the hydrogen atom. It is clear from this example how to proceed in the case that the observables are defined on overlapping s-tuples of successive states with $s > 2$, see also the introduction to Chapter 2.

5.1.2 Dimension Reduction

A solution to the practical problem of performing statistical tests for chain models with a huge state space by means of a generalized ϕ-divergence has been introduced in Example 3.6 of Section 3.2.2 already. The idea was the use of a linear mapping to reduce the dimension of the counter vector.

Consider again our example of the hydrogen atom and suppose that our spectrometer is a rather weak device such that we can only determine the color of a photon up to, say, $\{red, green, blue\} \cup \{black\}$[1]. Abbreviate the colors by their initial letter and denote the event *black*, which corresponds to a transition of the form (i, i), $i \in S$, by *bl*. Such a spectrometer defines a partition $\{\mathcal{S}_r, \mathcal{S}_g, \mathcal{S}_b, \mathcal{S}_{bl}\}$ of the set of possible transitions ("colors") $\overline{S}$. By (3.6) this induces a mapping

$$M = (m_{ij})_{(i,j)\in\overline{S}\times\{r,g,b,bl\}}, \qquad m_{ij} := \mathbf{1}_{\mathcal{S}_j}(i)$$

from $\overline{S}$ to $\{r, g, b, bl\}$. A statistical test on $\tilde{C}^{(n)} := \overline{C}^{(n)}M$ can now be performed by calculating the 4×4 covariance matrix $M'\overline{V}M$, its rank, and a weak inverse. Due to the dimension reduction and the fact that the absolute frequencies in $\tilde{C}^{(n)}$ sum up to n, the resulting statistic will have an asymptotic chi-square distribution with at most 3 degrees of freedom.

Although the above example still lacks physical significance, we have identified two important strategies – namely the use of overlapping chains and the use of linear mappings – which will be essential in the following more serious example.

5.2 Testing Long Period Generators

Overlapping Serial tests have been used extensively to analyze the quality of linear generators, see e.g. [30] and [56, 11]. The pertinence of such tests

[1] Here we definitely abandon the physical model where the wave lengths of the photons are mainly in the ultra-red band.

has been described in the introduction to Chapter 3. An important feature is their ability to distinguish between inversive and linear generators. Depending on the number m of sets in the partition $\mathcal{B} = \{B_1, \ldots, B_m\}$ of $[0,1)$, the test is feasible only in quite moderate dimensions s because of the large amount of memory that is needed to store the m^s components of the counter vector. This limits the applicability of the test to generators with "small" period length around 2^{32} to 2^{48}. Beyond this magic bound, almost any generator without serious defects passes the test. In this section we therefore consider a more stringent variant of the overlapping Serial test in order to be able to analyze two generators with a period length equal to $2^{800} - 1$.

As has already been discussed in Section 2.2, the overlapping chain allows to analyze transitions in the original chain. This can be important if one wants to detect hidden correlations in a process $(X_l)_{l \in \mathbb{N}_0}$. We will analyze two famous long-period generators[2], namely $T800$ [33] and $TT800$ [34]. It will be shown how theoretically known correlations in $T800$ (which originally led to the definition of $TT800$) can be revealed by an empirical test. The standard overlapping Serial test setup does not apply here since the correlations can be detected in high dimensions only. The dimension reduction technique will provide a test which is highly stringent on the one hand and can be interpreted as playing a simple gambling strategy in a fair game based on coin-flipping on the other hand. The latter property guarantees practical relevance of the improvements in $TT800$.

Both generators belong to the family of so-called twisted generalized feedback shift register generators (TGFSRs). Let ω, u and v be positive integers with $u > v$, and let A an $\omega \times \omega$ matrix with entries in $Z_2 = \{0,1\}^\omega$. We call b an ω-bit unsigned integer if $b = (c_0, \ldots, c_{w-1}) \in Z_2^\omega$. If an integer pseudorandom number $\texttt{prn} \in \mathbb{N}_0$ satisfies

$$\texttt{prn} = \sum_{i=0}^{\omega-1} c_i 2^i, \tag{5.1}$$

we can view b as a binary extension of $\texttt{prn}$. Here we call $c_{\omega-1}, c_{\omega-2}, \ldots, c_{\omega-k}$ the k most significant bits of $\texttt{prn}$.

A sequence $b_0, b_1, b_2, \ldots$ of ω-bit unsigned integers written as elements of Z_2 is a twisted generalized feedback shift register sequence with parameters

[2]We would like to thank Makoto Matsumoto from the Kyoto University in Japan for the many suggestions and assistance with respect to testing his generators.

(w, u, v, A) and initial values $b_{-u}, b_{-u+1}, \ldots, b_{-1}$ if it satisfies the recursion

$$b_l \equiv b_{l-u+v} + b_{l-u}A \pmod 2$$

for $l \geq 0$. The pseudorandom numbers $\texttt{prn}_l \in \{0, 1, \ldots, 2^w - 1\}$ are obtained by viewing b_l again as a block of binary digits which gives the finite binary extension of $\texttt{prn}_l$ for $l \geq 0$ as in (5.1). Normalized pseudorandom numbers $y_l \in [0, 1)$ are obtained[3] by setting $y_l := \texttt{prn}_l / 2^w$. In the case of $T800$ and $TT800$, $(\omega, u, v) = (32, 25, 7)$, and the starting values $b_{-25}, \ldots, b_{-1}$ and the matrix A are chosen as in [33, p. 262] and [34, p. 265]. The period length ϱ of both generators equals $2^{800} - 1$.

The difference between $T800$ and $TT800$ is that $T800$ belongs to the class of "rational-form" TGFSRs which means that the matrix A has a certain structure. This structure allows high-speed implementations of the generator but the resulting sequence suffers from correlations. $TT800$ thus uses so-called tempering, a simple but highly efficient additional transformation which is equivalent to a modification of the matrix A. To measure the impact of tempering on the quality of the generator, the authors consider the order of equidistribution to q-bit accuracy in [34]: let $q \in \mathbb{N}_{\omega u}$ and denote by $\lfloor \texttt{prn}_l \rfloor_2$ the element of the set $\{0, 1, 2, 3\}$ represented by the binary extension of the two most significant bits of $\texttt{prn}_l$. If e is the largest integer with the property that for every $k \in \mathbb{N}_e$ the sequence of vectors

$$(\lfloor \texttt{prn}_l \rfloor_2, \lfloor \texttt{prn}_{l+1} \rfloor_2, \ldots, \lfloor \texttt{prn}_{l+k-1} \rfloor_2)_{l \in \{0, \ldots, \varrho - 1\}} \in \{0, 1, 2, 3\}^k$$

contains each vector in $\{0, 1, 2, 3\}^k$ the same number of times except for the zero vector which appears once less often, we say, that the sequence $(\texttt{prn}_l)_{l \in \{0, \ldots, \varrho - 1\}}$ has an order of equidistribution to 2-bit accuracy equal to e.

The two most significant bits of $TT800$ have an order of equidistribution to 2-bit accuracy equal to 400 for instance. $TT800$ thus attains the trivial upper bound on the order of equidistribution to 2-bit accuracy which is feasible for generators with a period of $2^{800} - 1$. In sharp contrast, the two most significant bits of $T800$ have an order of equidistribution equal to 25 only, although the period length of the generator is the same.

In a first attempt to empirically reveal this defect of $T800$ we could try a standard overlapping Serial test with dimension $s = 26$ on the two most

[3] In the original code, the normalization $y_l := \texttt{prn}_l / (2^w - 1)$ is used, such that $y_l \in [0, 1]$.

significant bits of the generator. With respect to the test design in Section 3.1.3, this amounts to define the partition $\mathcal{B} = \{B_1, \ldots, B_4\}$, $B_i = [\frac{i-1}{4}, \frac{i}{4})$, $i \in \{1, 2, 3, 4\}$, of $[0, 1)$ and to count the number of points $(y_l, y_{l+1}, \ldots, y_{l+25})$, $0 \leq l \leq n - 1$, where n is the sample size, of the generator in every set of the partition $\mathcal{B}^{26}$ of $[0, 1)^{26}$. This partition contains $2^{52} = 4503599627370496$ sets. We would need a huge amount of RAM (≈ 4194304 Gigabyte) to maintain an array of counters in the internal memory of the computer running the test. The resulting test statistic would have an asymptotic chi-square distribution with $4^{26} - 4^{25} = 3377699720527872$ degrees of freedom. We did not try the standard overlapping Serial test in dimension 26 due to the corresponding computational difficulties. Results with a similar test setup in dimensions $s \in \{1, 2, 3, 4, 5\}$ did not show any defects of $T800$.

5.2.1 The Gambling Test

So as to be able to empirically reveal the known defects of $TG800$, we choose a different test setup and consider the notion of a fair game based on coin tossing. Each toss results in either "head" (h) or "tail" (t). Assume a gambler who keeps track of the last 52 outcomes and bets on h if h had occurred at least 26 times during the last 52 coin tosses. In this case, the next coin toss decides whether he wins (h occurs) or looses (t occurs). In the case that there have been less than 26 h during the last 52 coin tosses, the gambler skips the round and does not bet. Such kind of "optimistic" strategy relies on the assumption that the coin tends to cluster outcomes of a certain kind. In the following we construct an empirical test for pseudorandom numbers based on a simulation of this game. We call this test the *Gambling Test for Pseudorandom Number Generators.*

Denote the three possible outcomes of every game by 1 (win), 0 (not played), and -1 (loss). We use the independent chain with respect to $\mathbf{P} = (\frac{1}{2}, \frac{1}{2})$ and state space $S = \{h, t\}$ as model for the fair coin. For the overlapping chain $(\overline{S}, \overline{\mathbb{P}})$ with dimension $s = 53$ of $(S, \mathbb{P})$, $\#\overline{S} = 2^{53}$ and calculating the asymptotic expectation and covariance matrix of the vector $\overline{C}^{(n)}$ based on the formulas (2.11), (2.12) and (2.15) is quite expendable. Below we will see, that we can avoid the calculation of all these entities, however. $(S, \mathbb{P})$ being irreducible and aperiodic, the same is true for $(\overline{S}, \overline{\mathbb{P}})$ and the Central Limit Theorem applies.

We now apply the dimension reduction technique by grouping the possible states of the chain $(\overline{S}, \overline{\mathbb{P}})$ with respect to the three possible outcomes of each

game. We partition $\overline{S}$ into the sets $\mathcal{B}_1$, $\mathcal{B}_0$, and $\mathcal{B}_{-1}$ in the following way: $\mathcal{B}_0$ contains all states $(i_1, \ldots, i_{53}) \in \overline{S}$ where at most 25 of the i_k, $1 \le k \le 52$ equal h ("head"). All other states belong to either $\mathcal{B}_1$ if i_{53} equals h (win), or to $\mathcal{B}_{-1}$ if i_{53} equals t (loss). Let M be the $2^{53} \times 3$ matrix defined by

$$M = (m_{ij})_{(i,j) \in \mathbb{N}_{2^{53}} \times \mathbb{N}_3}, \quad m_{ij} = \mathbf{1}_{\mathcal{B}_j}(i). \tag{5.2}$$

This defines a mapping M from $\overline{S}$ to $\{1, 0, -1\}$. The counter $\tilde{C}^{(n)} = (\tilde{C}_1^{(n)}, \tilde{C}_0^{(n)}, \tilde{C}_{-1}^{(n)})$ defined by $\tilde{C}^{(n)} = \overline{C}^{(n)} M$ gives the number of wins, skipped games, and losses. By elementary combinatorics it is possible to find the asymptotic expectation $\tilde{\mathbf{P}} = \lim_{n\to\infty} \frac{1}{n} E[\tilde{C}^{(n)}]$ and covariance matrix $\tilde{V} = \lim_{n\to\infty} \frac{1}{n} V[\tilde{C}^{(n)}]$ of $\tilde{C}^{(n)}$, see Section 6.3 in the Appendix. In our case we have $\tilde{\mathbf{P}} \approx (0.277529, 0.444942, 0.277529)$. The rank $R(\tilde{V})$ equals 2 and a weak inverse $\tilde{V}^-$ of $\tilde{V}$ is given by

$$\tilde{V}^- \approx \begin{pmatrix} 2.908145042372 & 0.306521639158 & -3.21466668153 \\ 0.306521639158 & 0.084909927118 & -0.39143156628 \\ -3.214666681530 & -0.391431566276 & 3.60609824781 \end{pmatrix}.$$

The Gambling Test for (the sequence of) pseudorandom numbers $(\mathtt{prn}_l)_{l \ge 0}$ can be performed as follows: let $c_{\omega-1}^{(l)}$ and $c_{\omega-2}^{(l)}$ be the two most significant bits of $\mathtt{prn}_l$ and set

$$x_{2k} = \begin{cases} h & \text{if} \quad c_{\omega-1}^{(k)} = 1 \\ t & \text{if} \quad c_{\omega-1}^{(k)} = 0 \end{cases} \quad \text{and} \quad x_{2k+1} = \begin{cases} h & \text{if} \quad c_{\omega-2}^{(k)} = 1 \\ t & \text{if} \quad c_{\omega-2}^{(k)} = 0 \end{cases}, \quad k \ge 0.$$

By the null-hypotheses that the $\mathtt{prn}_l$ are realizations of independent random variables distributed uniformly on $\{0, \ldots, \varrho - 1\}$, we may interpret the sequence $(x_l)_{l \ge 0}$ as a sequence of independent fair coin tosses. Let $n \in \mathbb{N}$ be an arbitrary but fixed sample size and calculate a realization of the counter $\tilde{C}^{(n)}$ based on the coin tosses $x_0, \ldots, x_{n+51}$, where $x_0, \ldots, x_{51}$ are needed to initialize the memory of the gambling strategy. The first possibility to bet arises after the 52'nd coin toss. We again use the generalized log-likelihood divergence of Section 5.1. With $\hat{\mathbf{P}}^{(n)} = \frac{1}{n} \tilde{C}^{(n)}$, the statistic

$$\tilde{I} = \tilde{I}_{\tilde{V}^-, \varphi}(\hat{\mathbf{P}}^{(n)}, \tilde{\mathbf{P}}) = 2n \sum_{(i,j) \in \{1,0,-1\}^2} \tilde{V}_{ij}^- \tilde{\mathbf{P}}_i \tilde{\mathbf{P}}_j \phi^{\varphi} \left(\frac{\hat{P}_i^{(n)}}{\tilde{\mathbf{P}}_i}, \frac{\hat{P}_j^{(n)}}{\tilde{\mathbf{P}}_j} \right)$$

is for large n approximately chi-square distributed with 2 degrees of freedom. In the above, $\tilde{I}$ has been calculated on the basis of the $\tilde{n} = \lceil \frac{n+51}{2} \rceil + 1$ PRNs $\mathtt{prn}_0, \ldots, \mathtt{prn}_{\tilde{n}-1}$.

Instead of just computing a single value $\tilde{I}$ and performing the test by rating this value with respect to a suitable rejection region, we replicate the test 64 times by computing the values $\tilde{I}_0, \ldots, \tilde{I}_{63}$ where $\tilde{I}_i$ is based on the PRNs $\mathtt{prn}_{i\tilde{n}}, \ldots, \mathtt{prn}_{i\tilde{n}+\tilde{n}-1}$. The empirical distribution function

$$\hat{F}_{64}(t) = \frac{1}{64} \# \left\{ \tilde{I}_i \leq t \ : \ i \in \{0, \ldots, 63\} \right\}, \quad t \in \mathbb{R}$$

will be compared to the distribution function $F(t)$ of a chi-square distributed random variable with 2 degrees of freedom by means of a two-sided Kolmogorov-Smirnov test.

5.2.2 Empirical Results

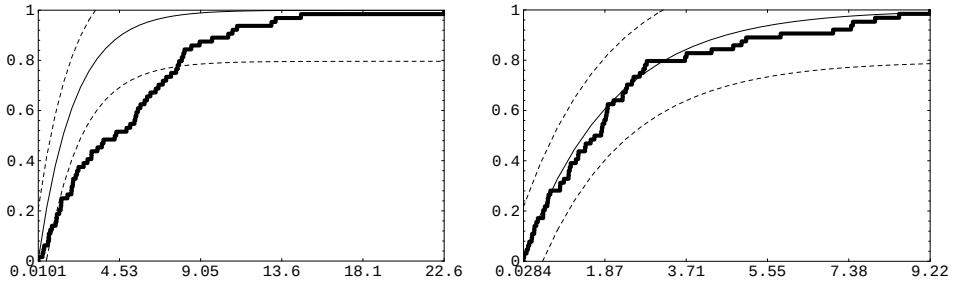
Figure 5.1 shows the empirical distribution function (bold line) $\hat{F}_{64}(t)$ of $\tilde{I}_0, \ldots, \tilde{I}_{63}$ and the (asymptotic) theoretical distribution function (thin line) $F(t)$ of $\tilde{I}_{\tilde{V}^-,\varphi}(\hat{\mathbf{P}}^{(n)}, \tilde{\mathbf{P}})$ for the sample size $n = 2^{22}$. Following [17, p. 183], the rejection region of the two-sided Kolmogorov-Smirnov test with 64 samples and with a level of significance of 0.01 is approximately $[1.63, \infty)$. The two dashed lines $(F(t) \pm 1.63 \, \frac{1}{\sqrt{64}})$ give a 99% confidence interval for the empirical distribution function based on this rejection region. A generator is rejected by the test if the empirical distribution function does not lie completely inside these boundaries. The whole test uses about 2^{27} pseudo-random numbers which is – in the case of $T800$ and $TT800$ – a negligible fraction of the whole period length.

The values of the two-sided Kolmogorov-Smirnov test statistic are 3.19 for $T800$ and 0.723 for $TT800$. The test clearly rejects $T800$ and empirically confirms the theoretically foreseen improvement caused by using tempering in $TT800$.

The generalized log-likelihood divergence is a member of the family of generalized ϕ^{φ}-divergences $\tilde{I}_{\overline{\Sigma},\varphi}$ with functions φ in the class φ_α introduced by Liese and Vajda [31], see the introduction to Chapter 4. There we had

$$\varphi_\alpha(u) = \begin{cases} u - 1 - \ln u & \alpha = 0 \\ \frac{\alpha u + 1 - \alpha - u^\alpha}{\alpha(1-\alpha)} & \alpha \in \mathbb{R} \setminus \{0, 1\} \\ 1 - u + u \ln u & \alpha = 1 \end{cases}$$

The case $\alpha = 1$ yields the log-likelihood divergence I_{φ_1} and – owing to Theorem 4.2 – the generalized log-likelihood divergence $\tilde{I}_{\tilde{V}^-,\varphi_1}$. One might



$T800$ $TT800$

Figure 5.1: Results of the Gambling Test for Pseudorandom Number Generators $T800$ and $TT800$, given in terms of the empirical distribution function (bold line) and a 99% confidence band.

ask, whether this special selection of α has a strong influence on the test results. We cannot give a thorough answer in this thesis. Instead, we supply two plots in Figure 5.2 which summarize some aspects of the behaviour of different generalized ϕ^{φ_α}-divergences where we vary the parameter $\alpha \in \{-16, -15.5, -15, \ldots, 15.5, 16\}$ and the sample size $n \in \{2^3, \ldots, 2^{24}\}$.

For each pair (n, α) we calculate a realization of the counter vector $\tilde{C}^{(n)}$ based on $\tilde{n}$ PRNs. Here we use consecutive non-overlapping $\tilde{n}$-tuples of PRNs from the generator for each possible combination (n, α). Again, let $\hat{\mathbf{P}}^{(n)} = \frac{1}{n}\tilde{C}^{(n)}$ and denote by $\tilde{I}_{n,\alpha}$ the statistic

$$\tilde{I}_{n,\alpha} = \tilde{I}_{\tilde{V}^-,\varphi_\alpha}(\hat{\mathbf{P}}^{(n)}, \tilde{\mathbf{P}}) = 2n \sum_{(i,j)\in\{1,0,-1\}^2} \tilde{V}_{ij}^- \tilde{\mathbf{P}}_i \tilde{\mathbf{P}}_j \phi^{\varphi_\alpha}\left(\frac{\hat{P}_i^{(n)}}{\tilde{\mathbf{P}}_i}, \frac{\hat{P}_j^{(n)}}{\tilde{\mathbf{P}}_j}\right).$$

For each realization of the counter vector $\tilde{C}^{(n)}$ we now compute the value $\tilde{I}_{n,\alpha}$ of the according generalized ϕ^{φ_α}-divergence. The grey shade of the according pixel in the plots corresponds to the upper tail probability $U_{n,\alpha} = P[\chi_2^2 > \tilde{I}_{n,\alpha}]$, where χ_2^2 denotes a chi-square random variable with 2 degrees of freedom, and $\tilde{I}_{n,\alpha}$ denotes a realization of the test statistic with the same name. Under the null hypothesis, $U_{n,\alpha}$ is uniformly distributed on the unit interval. Values near 0 are plotted as almost black pixels and indicate that the value of the test statistic is large, i.e. the weighted distance between

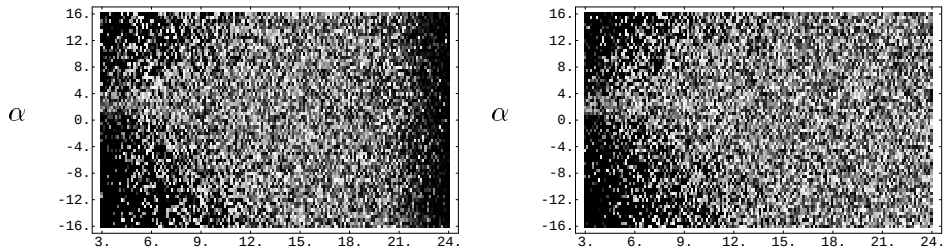


Figure 5.2: Results of the Gambling Test for Pseudorandom Number Generators $T800$ and $TT800$, given in terms of the upper tail probability. Darker colors indicate higher values of the test statistic

the counter vector and its expectation is large. This situation occurs either for small sample sizes n and values α far from 2 (recall, that $\tilde{I}_{\tilde{V}^-,\varphi_2}$ is the quadratic form in the weak inverse), or for sample sizes $n \geq 2^{22}$ in the case of $T800$. The former indicates the approximation error stemming from the fact that we calculate $U_{n,\alpha}$ on the basis of the asymptotic chi-square distribution and appears for both generators. The latter is due to the aforementioned correlations in the pseudorandom numbers obtained from $T800$ and does not appear in the results of $TT800$. Almost white pixels indicate small values of $\tilde{I}_{n,\alpha}$.

The interesting observation from Figure 5.2 is that all members of the generalized ϕ^{φ_α}-divergence tend to reject $T800$ for sample sizes n equal or greater 2^{22}. In the given test design, no test statistic outperforms the others with respect to the detection of $T800$'s correlation structure.

In summary, the Gambling Test for pseudorandom number generators has proven to be able to detect correlations in huge period linear generators. It seems that the state space of $T800$ being 25 times 32-bit words with excellent equidistribution properties up to dimension 25, any empirical attack of such a generator will require a memory size of order 26 or more in order to be "successful". Here we consider an empirical test "successful" if it is able to reveal known defects of a generator.

The multiple-recursive matrix method described in [41, Section 4.1 and 5.2] gives an unified framework for linear generation methods for pseudorandom numbers. It includes the multiplicative linear congruential method with prime modulus, the multiple-recursive congruential method with prime modulus, the generalized feedback shift register (GFSR) method, and the TGFSR generator as special cases. From the implementation point of view (i.e. for the reason of speed), this method is interesting only if one manages to achieve a representation of a specific generator where the calculation of the next pseudorandom number uses linear combinations of only a few bits of the preceding pseudorandom numbers. Such a case occurs, for example, if the defining matrices ($A_0, \ldots, A_{k-1}$ in [41]) are sparse (i.e. contain only few non-zero elements). In a forthcoming paper with Makoto Matsumoto we apply the Gambling Test to several well-known "sparse" generators and show the potential risks of such high-speed algorithms. Speaking loosely, the gambling strategy works well since states with many zero-bits are more likely to be followed by states with many zero-bits.

Chapter 6

Appendix

6.1 Lemmata

We first recall several properties and definitions concerning real valued random variables. Here and in the sequel we shall denote the distribution function of a random variable U by F_U, where $F_U(t) = P[U \leq t]$. $(U_n)_{n\in\mathbb{N}}$ and $(V_n)_{n\in\mathbb{N}}$ will denote sequences of real random variables.

The proof of the following obvious statement is given for the reason of completeness, only.

Lemma 6.1 *If U is a real valued random variable and if $\epsilon > 0$, then there exists a $x \in \mathbb{R}$ such that $P[|U| \geq x] < \epsilon$.*

Proof: For $n \in \mathbb{N}$ let $A_n = [-n, n]\backslash(-n+1, n-1)$. Since $\bigcup_{n\in\mathbb{N}} A_n = \mathbb{R}$, the probability $P[U \in \bigcup_{n\in\mathbb{N}} A_n]$ equals 1. Because the A_n are pairwise disjoint, we have $1 = \sum_{n\in\mathbb{N}} P[U \in A_n]$. Thus there exists a $x \in \mathbb{N}$ such that $\sum_{n=1}^{x-1} P[U \in A_n] \geq 1 - \epsilon$. ∎

Definition 6.1 (Convergence of Random Variables) *A sequence of random variables $(U_n)_{n\in\mathbb{N}}$ converges in distribution to a random variable U, $U_n \xrightarrow{\mathrm{d}} U$ if $F_{U_n}(t) \to F_U(t)$ as $n \to \infty$ for all t, where $F_U(t)$ is continuous. This condition is also called 'weak convergence of the distribution functions'.*
We say that U_n converges in probability to U, $U_n \xrightarrow{\mathrm{p}} U$ if for all $\epsilon > 0$: $P[|U_n - U| > \epsilon] \to 0$

In general, convergence in probability implies convergence in distribution. In the special case that U is constant almost surely (a.s.), the both types of convergence are equivalent: let $U_n \xrightarrow{\mathrm{d}} U$, where $P[U = u] = 1$ for a constant $u \in \mathbb{R}$. Then F_U is continuous in every point $t \neq u$. Thus for every $\epsilon > 0$

$$P[|U_n - U| > \epsilon] = P[|U_n - u| > \epsilon] \leq 1 - F_{U_n}(u + \epsilon) + F_{U_n}(u - \epsilon),$$

where $F_{U_n}(u + \epsilon) \to F_U(u + \epsilon) = 1$ and $F_{U_n}(u - \epsilon) \to F_U(u - \epsilon) = 0$, respectively. Note that all variables U_n and U itself have to be defined on the same probability space in order that the concept of convergence in probability may be applied. However, the relation $U_n \xrightarrow{\mathrm{d}} U$ makes sense also if the U_n are defined on different spaces.

We further recall that convergence in distribution implies convergence of the form $P[U_n \in \mathcal{A}] \to P[U \in \mathcal{A}]$, as long as the boundary of the measurable set $\mathcal{A}$ is a null-set with respect to $P \circ U$. Thus $U_n \xrightarrow{\mathrm{d}} U$ implies

$$P[|U_n| \geq x] \to P[|U| \geq x], \tag{6.1}$$

provided that $P[|U| = x] = 0$. Closely connected to convergence in probability is the notion of tightness:

Definition 6.2 (Tightness) *A sequence of real valued random variables $(U_n)_{n \in \mathbb{N}}$ is called 'tight', if for all $\epsilon > 0$ there exists a $x > 0$ such that for all $n \in \mathbb{N}$: $P[|U_n| \geq x] < \epsilon$.*

Tightness guarantees the existence of weakly converging subsequences of arbitrary subsequences of U_n, where each such limit is a random variable itself. So, if $U_n \xrightarrow{\mathrm{d}} U$, then U_n is necessarily tight. This can also be seen from the following Lemma.

Lemma 6.2 *If $U_n \xrightarrow{\mathrm{d}} U$, $V_n \xrightarrow{\mathrm{d}} V$, where U and V are real valued, and $\mu \in \mathbb{R}$ is constant, then μU_n and $\mu U_n V_n$ are tight.*

Proof: The case $\mu = 0$ is trivial. Now assume w.l.o.g. that $\mu > 0$. For every $\epsilon > 0$ there exists a x_0 such that $P[\mu|U| \geq x_0] < \frac{\epsilon}{2}$ and that $P[\mu|U| = x_0] = 0$. This follows from Lemma 6.1 and the fact that every real random variable allows utmost countable many values z for which $P[U = z] > 0$. Applying (6.1), we have a $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$|P[\mu|U_n| \geq x_0] - P[\mu|U| \geq x_0]| < \frac{\epsilon}{2}$. Thus, for all $n \geq n_0$, $P[\mu|U_n| \geq x_0] < \epsilon$. Further, each variable U_k, $1 \leq k < n_0$ fulfills $P[\mu|U_k| \geq x_k] < \epsilon$ for some $x_k \in \mathbb{R}$. Taking x the maximum of x_0 and the x_k completes the proof of the first statement. As to the second statement, we again assume $\mu > 0$. Now for every $\epsilon > 0$, we have a x_0 such that $P[\sqrt{\mu}|U| \geq x_0] < \frac{\epsilon}{4}$, $P[\sqrt{\mu}|V| \geq x_0] < \frac{\epsilon}{4}$, $P[|U| = \sqrt{\frac{x_0}{\mu}}\,] = 0$, and $P[|V| = \sqrt{\frac{x_0}{\mu}}\,] = 0$. For the same ϵ we also have a $n_0 \in \mathbb{N}$ such that for every $n \geq n_0$

$$\begin{aligned} P\left[\mu|U_nV_n| \geq x_0\right] &\leq P\left[\left(|U_n| \geq \sqrt{\frac{x_0}{\mu}}\right) \vee \left(|V_n| \geq \sqrt{\frac{x_0}{\mu}}\right)\right] \leq \\ &\leq P\left[|U_n| \geq \sqrt{\frac{x_0}{\mu}}\right] + P\left[|V_n| \geq \sqrt{\frac{x_0}{\mu}}\right] \leq \\ &\leq P\left[|U| \geq \sqrt{\frac{x_0}{\mu}}\right] + P\left[|V| \geq \sqrt{\frac{x_0}{\mu}}\right] + \frac{\epsilon}{4} + \frac{\epsilon}{4}, \end{aligned}$$

where the last line follows by applying (6.1) two times. Thus $P\left[\mu|U_nV_n| \geq x_0\right] \leq \epsilon$ for all $n \geq n_0$. The tightness of the sequence μU_nV_n now follows readily. ■

The following technical Lemma will be used below.

Lemma 6.3 *If $(U_n)_{n\in\mathbb{N}}$ and $(V_n)_{n\in\mathbb{N}}$ satisfy $U_n \xrightarrow{\mathrm{p}} 0$ and $V_n \xrightarrow{\mathrm{p}} 0$ then for all $\alpha, \beta > 0$*

$$P[|U_n| < \alpha \wedge |V_n| < \beta] \to 1, \text{ as } n \to \infty.$$

Proof: For arbitrary $\epsilon > 0$ there exists a $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $P[|U_n| \geq \alpha] < \frac{\epsilon}{2}$ and $P[|V_n| \geq \beta] < \frac{\epsilon}{2}$ and consequently

$$P[|U_n| < \alpha \wedge |V_n| < \beta] = 1 - P[|U_n| \geq \alpha \vee |V_n| \geq \beta] \geq$$

$$\geq 1 - (P[|U_n| \geq \alpha] + P[|V_n| \geq \beta]) \geq 1 - \epsilon.$$

■

We now establish a relation between the rate of growth of the normalizing constants in a convergence-in-distribution statement and the distance of the random variables to their expectation.

Lemma 6.4 *Let U_n be a sequence of random variables, $u \in \mathbb{R}$ be fixed, and let $(c_n)_{n\in\mathbb{N}}$ be a sequence of positive numbers diverging to infinity. If $c_n\,(U_n - u) \xrightarrow{\mathrm{d}} U$ for some real valued random variable U, then $|U_n - u| \xrightarrow{\mathrm{p}} 0$.*

Proof: Let $\epsilon > 0$ and $\eta > 0$ be arbitrary but fixed. Choose $x > 0$ such that $P[|U| > x] < \frac{\epsilon}{2}$ and $P[|U| = x] = 0$. By (6.1) we have a $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$: $P[c_n|U_n - u| \geq x] \leq P[|U| \geq x] + \frac{\epsilon}{2}$. Since $c_n \to \infty$, there also exists a n_1, such that for all $n \geq n_1$, $\frac{x}{c_n} < \eta$. Thus for all $n \geq \max\{n_0, n_1\}$,

$$P\left[|U_n - u| \geq \eta\right] \leq P\left[c_n|U_n - u| \geq x\right] \leq P\left[|U| > x\right] + \frac{\epsilon}{2} \leq \epsilon.$$

■

The following Corollary is essential in the proof of the asymptotic distribution of the generalized ϕ-divergence.

Corollary 6.5 *Let $\hat{\mathbf{P}}^{(n)}$ be a sequence of random vectors in $\mathbb{R}^m$ and $\mathbf{P} \in D_m$ such that $\sqrt{n}(\hat{\mathbf{P}}^{(n)} - \mathbf{P}) \xrightarrow{\mathrm{d}} \mathcal{N}(\emptyset, \Sigma)$, where Σ is a covariance matrix, then $|\hat{P}_i^{(n)} - P_i| \xrightarrow{\mathrm{p}} 0$ for every $i \in \{1, \ldots, m\}$.*

Proof: Convergence in R^m implies coordinate-wise convergence, such that for each $i \in \{1, \ldots, m\}$ we have $\sqrt{n}(\hat{P}_i^{(n)} - P_i) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma_{ii})$. Now apply Lemma 6.4. ■

An often used technical device within the scope of convergence in distribution analysis is the Theorem of Slutsky: assume that a sequence X_n of random variables can be written $X_n = Y_n + U_n V_n + W_n$, where the limiting distribution of Y_n is known, $Y_n \xrightarrow{\mathrm{d}} Y$. Provided that $U_n \xrightarrow{\mathrm{d}} U$, $V_n \xrightarrow{\mathrm{p}} 0$ and $W_n \xrightarrow{\mathrm{p}} 0$, the theorem allows to conclude $U_n V_n + W_n \xrightarrow{\mathrm{p}} 0$. In this case, X_n and Y_n have the same asymptotic distribution Y, see e.g.[3, Theorem 25.4]. In the next lemma, we generalize parts of the Theorem of Slutsky to the case that U_n is tight.

Lemma 6.6 *If U_n is a tight sequence of random variables and $V_n \xrightarrow{\mathrm{p}} 0$ and $W_n \xrightarrow{\mathrm{p}} 0$ then*
i) $U_n V_n \xrightarrow{\mathrm{p}} 0$, and
ii) $V_n + W_n \xrightarrow{\mathrm{p}} 0$

Proof: As for i), let $\epsilon > 0$ and $\eta > 0$ arbitrary but fixed. We have to show that there exists a $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$: $P[|V_n U_n| \geq \epsilon] < \eta$. For every $x > 0$ we have

$$P[|V_n U_n| \geq \epsilon] \quad = \quad P[|V_n U_n| \geq \epsilon \wedge |V_n| < \frac{\epsilon}{x}] + P[|V_n U_n| \geq \epsilon \wedge |V_n| \geq \frac{\epsilon}{x}]$$

$$\begin{aligned} &\leq P[\epsilon \leq |U_n V_n| \leq |U_n|\frac{\epsilon}{x}] + P[|V_n| \geq \frac{\epsilon}{x}] \\ &\leq P[x \leq |U_n|] + P[|V_n| \geq \frac{\epsilon}{x}] \end{aligned}$$

By the tightness of U_n there exists a x such that $P[x \leq |U_n|] < \frac{\eta}{2}$ for all $n \in \mathbb{N}$. Also, by the convergence of V_n there exists a n_0 such that for all $n \geq n_0$, $P[|V_n| \geq \frac{\epsilon}{x}] < \frac{\eta}{2}$. As for ii), we note that owing to the triangle-inequality $P[|V_n + W_n| < \epsilon] \geq P[|V_n| < \frac{\epsilon}{2} \wedge |W_n| < \frac{\epsilon}{2}]$ and apply Lemma 6.3 with $\alpha = \beta = \frac{\epsilon}{2}$.

■

Corollary 6.7 *If* $(\mu_n)_{n\in\mathbb{N}} \to \mu \in \mathbb{R}$ *and* $V_n \xrightarrow{\mathrm{p}} 0$, *then also* $\mu_n V_n \xrightarrow{\mathrm{p}} 0$.

The Corollary follows from the fact that we may choose $U_n = \mu_n$ which is a sequence of constant random variables that trivially fulfills the condition of Lemma 6.6.

We finish this section by a Lemma on multivariate normal distributions which states that a linear transform of a multivariate normal distribution is multivariate normal again.

Lemma 6.8 *Let* μ *in* $\mathbb{R}^n$, $n \in \mathbb{N}$ *arbitrary but fixed,* Σ *a* $n \times n$ *positive semidefinite symmetric real matrix, and let* $X \sim \mathcal{N}(\mu, \Sigma)$ *be distributed multivariate normal with mean* μ *and covariance matrix* Σ. *Further assume a* $n \times k$ *real matrix* M, $k \in \mathbb{N}$, *and put* $Y = XM$. *Then* Y *is distributed multivariate normal* $Y \sim \mathcal{N}(\mu M, M'\Sigma M)$ *with mean* μM *and covariance matrix* $M'\Sigma M$.

For a simple proof see e.g. [3, p. 398].

6.2 Proof of the Asymptotic Distribution of $I_{\overline{\Sigma},\phi}$

Here we show that under the assumptions of Theorem 4.1, both remainder terms, $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ and $U(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ converge in probability to zero so that the asymptotic distribution of $I_{\overline{\Sigma},\phi}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$ is equal to that of $T(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$.

As to the second term, recall that $U(\hat{\mathbf{P}}, \mathbf{P}) \neq 0$ implies that $\hat{\mathbf{P}} \notin K_\delta(\mathbf{P})$. By applying Lemma 6.3 repeatedly to Corollary 6.5 we get the convergence of

$\hat{\mathbf{P}}^{(n)}$ to $\mathbf{P}$ with respect to the maximum norm in $\mathbb{R}^m$ and consequently the convergence of $P[\hat{\mathbf{P}}^{(n)} \notin K_\delta(\mathbf{P})]$ to zero. Thus we have for every $\epsilon > 0$

$$P\left[\left|U(\hat{\mathbf{P}}^{(n)}, \mathbf{P})\right| \geq \epsilon\right] \leq P\left[\hat{\mathbf{P}}^{(n)} \notin K_\delta(\mathbf{P})\right] \to 0,$$

that is, $U(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\text{P}} 0$ as n goes to infinity.

In addition we have to show that the first remainder term, $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, converges in probability to zero. Recall that $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \frac{n}{2} \sum_{i,j=1}^{m} c_{ij}\, P_i P_j\, R_{ij}(\hat{\mathbf{P}}^{(n)}, \mathbf{P})$, where

$$R_{ij}(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) = \epsilon_{xx}^{(n)}(i,j) Q_i^2 + 2\epsilon_{xy}^{(n)}(i,j) Q_i Q_j + \epsilon_{yy}^{(n)}(i,j) Q_j^2.$$

Discarding constants by applying Corollary 6.7, we get $R(\hat{\mathbf{P}}^{(n)}, \mathbf{P}) \xrightarrow{\text{P}} 0$ provided that (i), (ii), and (iii) hold for all $i, j \in \{1, \ldots, m\}$:

(i) $\epsilon_{xx}^{(n)}(i,j)\ \ n\left(\hat{P}_i^{(n)} - P_i\right)^2 \xrightarrow{\text{P}} 0$,

(ii) $\epsilon_{xy}^{(n)}(i,j)\ \ n\left(\hat{P}_i^{(n)} - P_i\right)\left(\hat{P}_j^{(n)} - P_j\right) \xrightarrow{\text{P}} 0$, and

(iii) $\epsilon_{yy}^{(n)}(i,j)\ \ n\left(\hat{P}_j^{(n)} - P_j\right)^2 \xrightarrow{\text{P}} 0$.

We prove these by showing that each $\epsilon^{(n)}(i,j)$ converges in probability to zero, and that the remaining terms are tight sequences of random variables. Then we may apply Lemma 6.6 i) to the product.

Let i, j, and n be arbitrary but fixed and let $\epsilon > 0$. By the continuity of $\phi_{xy}(\cdot,\cdot)$ at $(1,1)$ there exist $\alpha = \alpha(\epsilon) > 0$ and $\beta = \beta(\epsilon) > 0$ such that $\max\{\alpha, \beta\} \leq \delta$, where $\delta > 0$ is given by the assumptions in Theorem 4.1, and

$$|a_{ij}^{(n)} - 1| < \alpha \wedge |b_{ij}^{(n)} - 1| < \beta \Rightarrow \left|\epsilon_{xy}^{(n)}(i,j)\right| < \epsilon. \tag{6.2}$$

Note that α and β do not depend on n. From (4.8) we have $|a_{ij}^{(n)} - 1| \leq |\frac{\hat{P}_i^{(n)}}{P_i} - 1|$ and $|b_{ij}^{(n)} - 1| \leq |\frac{\hat{P}_j^{(n)}}{P_j} - 1|$ such that applying (6.2) and (4.8) we get

$$\begin{aligned} P\left[\left|\epsilon_{xy}^{(n)}(i,j)\right| < \epsilon\right] &\geq P\left[|a_{ij}^{(n)} - 1| < \alpha \wedge |b_{ij}^{(n)} - 1| < \beta\right] \geq \\ &\geq P\left[|\frac{\hat{P}_i^{(n)}}{P_i} - 1| < \alpha \wedge |\frac{\hat{P}_j^{(n)}}{P_j} - 1| < \beta\right]. \end{aligned}$$

Now let n go to infinity. Applying Corollary 6.5 to the assumption $\sqrt{n}\left(\hat{\mathbf{P}}^{(n)} - \mathbf{P}\right) \xrightarrow{\text{d}} \mathcal{N}(\emptyset, \Sigma)$ in Theorem 4.1, we get $\hat{P}_i^{(n)} \xrightarrow{\text{P}} P_i$, for all

$i \in \{1, \ldots, m\}$ and consequently $\frac{\hat{P}_i^{(n)}}{P_i} - 1 \xrightarrow{\mathrm{P}} 0$ and $\frac{\hat{P}_j^{(n)}}{P_j} - 1 \xrightarrow{\mathrm{P}} 0$. Applying Lemma 6.3 yields $P\left[|\frac{\hat{P}_i^{(n)}}{P_i} - 1| < \alpha \wedge |\frac{\hat{P}_j^{(n)}}{P_j} - 1| < \beta\right] \to 1$ for all $\alpha, \beta > 0$. We have thus shown that for arbitrary i and j and for every $\epsilon > 0$,

$$P[|\epsilon_{xy}^{(n)}(i,j)| < \epsilon] \to 1 \text{ as } n \to \infty.$$

A similar calculation yields $\epsilon_{xx}^{(n)}(i,j) \xrightarrow{\mathrm{P}} 0$ and $\epsilon_{yy}^{(n)}(i,j) \xrightarrow{\mathrm{P}} 0$.

It remains to check the tightness of the remaining terms. For the sake of simplicity we let $I_n := \sqrt{\frac{n}{\Sigma_{ii}}}(\hat{P}_i^{(n)} - P_i)$ and $J_n := \sqrt{\frac{n}{\Sigma_{jj}}}(\hat{P}_j^{(n)} - P_j)$. Then clearly $I_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$, $J_n \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$.

As for (i) and (iii), we now apply Lemma 6.2 for $\mu = \sqrt{\Sigma_{ii}}$ and $U_n = I_n^2$, which is an asymptotically chi-square distributed sequence. This yields the desired tightness of $n\left(\hat{P}_i^{(n)} - P_i\right)^2$.

Regarding (ii), applying Lemma 6.2 to $\mu = \sqrt{\Sigma_{ii} \cdot \Sigma_{jj}}$, $U_n = I_n$, $V_n = J_n$ yields the tightness of $n\left(\hat{P}_i^{(n)} - P_i\right)\left(\hat{P}_j^{(n)} - P_j\right)$. ∎

6.3 Expectation and Covariance Matrix of $\tilde{C}^{(n)}$

In Section 5.2 of Chapter 5 we introduced the gambling test for pseudorandom number generators. Here we supplement the expectations and covariances of the counter $\tilde{C}^{(n)}$.

Let X_l, $l \in \mathbb{N}_0$, $X_l \in \{h,t\}$ denote the state of the chain $(S, \mathbb{P})$ with state space $\{h,t\}$ and transition matrix $\mathbb{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$ at time l and let $\mathbf{P}_0 = (1/2, 1/2)$ be the initial distribution. Since $(S, \mathbb{P})$ is independent with respect to $\mathbf{P} = (1/2, 1/2)$, X_l is an i.i.d. sequence of random variables distributed uniformly on $\{h,t\}^{\mathbb{N}_0}$. Recall that the sequence $(x_l)_{l \geq 0}$ in Section 5.2 denotes a realization of $(X_l)_{l \geq 0}$ based on the pseudorandom numbers $(\mathtt{prn}_l)_{l \geq 0}$. We put

$$Z_l := \begin{cases} 1 & \text{if } \sum_{j=l}^{l+51} \mathbf{1}_{\{h\}}(X_j) \geq 26 \text{ and } X_{l+52} = h \\ 0 & \text{if } \sum_{j=l}^{l+51} \mathbf{1}_{\{h\}}(X_j) \leq 25 \\ -1 & \text{if } \sum_{j=l}^{l+51} \mathbf{1}_{\{h\}}(X_j) \geq 26 \text{ and } X_{l+52} = t \end{cases}$$

the outcome of the l'th game. Recall, that the first 52 coin tosses are needed to initialize the memory of the gambling strategy. Since $(X_l)_{l\geq 0}$ is a stationary sequence, the same obviously holds for $(Z_l)_{l\geq 0}$.

The gambler being mainly interested in the expected win $E[Z_l]$ per game, we will go further and analyze the asymptotic expectation and covariance of the counters $\tilde{C}^{(n)} := (\tilde{C}_1^{(n)}, \tilde{C}_0^{(n)}, \tilde{C}_{-1}^{(n)})$, where $\tilde{C}_i^{(n)} = \#\{Z_l = i \ : \ 0 \leq l \leq n-1\}$, $i \in \{1, 0, -1\}$. Note, that the above definition of $\tilde{C}^{(n)}$ is equivalent to that given in Section 5.2, where we had $\tilde{C}^{(n)} = \overline{C}^{(n)} M$ for the matrix (5.2)

$$M = (m_{ij})_{(i,j)\in \mathbb{N}_{2^{53}} \times \mathbb{N}_3}, \quad m_{ij} = \mathbf{1}_{\mathcal{B}_j}(i).$$

As in Section 3.1.3, we slightly modify our model in the following. This simplifies the calculations significantly and has no effect on the expectations and an asymptotically negligible effect on the covariances: for an arbitrary but fixed sample size $n \in \mathbb{N}$, $n \geq 53$, let the sequence $(X_l^*)_{l\geq 0}$ be defined by $X_l^* = X_k$ with $k = l \pmod n$ and let Z_l be defined as above but by using X_l^* instead of X_l. This change affects only the last 52 outcomes $Z_{n-52}, \ldots, Z_{n-1}$. Note, that $(Z_l)_{0\leq l\leq n-1}$ still is a stationary sequence.

As to the expectation $\tilde{\mathbf{P}} = \frac{1}{n} E[\tilde{C}^{(n)}]$, $\tilde{\mathbf{P}} = (\tilde{P}_1, \tilde{P}_0, \tilde{P}_{-1})$, we have according to the linearity of the expectation and the stationarity of $(Z_l)_{0\leq l\leq n-1}$,

$$\tilde{P}_i = \frac{1}{n} E[\#\{Z_l = i \ : \ 0 \leq l \leq n-1\}] = \frac{1}{n} \sum_{i=0}^{n-1} E[\mathbf{1}_{\{i\}}(Z_l)] = P[Z_0 = i].$$

These probabilities calculate to

$$(\tilde{P}_1, \tilde{P}_0, \tilde{P}_{-1}) = (\frac{1-\tilde{P}_0}{2}, \tilde{P}_0, \frac{1-\tilde{P}_0}{2})$$

with $\tilde{P}_0 = \frac{1}{2}(1 - \binom{52}{26}/2^{52})$, where $\binom{52}{26}$ denotes the binomial coefficient, so that we get the following numerical approximation

$$(\tilde{P}_1, \tilde{P}_0, \tilde{P}_{-1}) \approx (0.277529, 0.444942, 0.277529).$$

As to the covariance matrix $\Sigma = (\sigma_{ij})_{(i,j)\in\{1,0,-1\}^2}$ we have

$$\sigma_{ij} = \frac{1}{n} Cov\left[\tilde{C}_i^{(n)}, \tilde{C}_j^{(n)}\right] = \frac{1}{n}\left(E[\tilde{C}_i^{(n)}\tilde{C}_j^{(n)}] - E[\tilde{C}_i^{(n)}]E[\tilde{C}_j^{(n)}]\right).$$

Since $\tilde{C}_i^{(n)} = \sum_{l=0}^{n-1} \mathbf{1}_{\{i\}}(Z_l)$ we get by the linearity of the expectation and by the stationarity of $(Z_l)_{0\le l\le n-1}$,

$$\begin{aligned} E[\tilde{C}_i^{(n)}\tilde{C}_j^{(n)}] &= \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} E[\mathbf{1}_{\{i\}}(Z_k)\mathbf{1}_{\{j\}}(Z_l)] \\ &= \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} P[Z_k = i \text{ and } Z_l = j] = \\ &= n\sum_{l=0}^{n-1} P[Z_0 = i \text{ and } Z_l = j]. \end{aligned}$$

Similarly,

$$\begin{aligned} E[\tilde{C}_i^{(n)}]E[\tilde{C}_j^{(n)}] &= \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} P[Z_k = i]P[Z_l = j] = \\ &= n\sum_{l=0}^{n-1} P[Z_0 = i]P[Z_l = j] = \\ (\quad &= n^2\tilde{\mathbf{P}}_i\tilde{\mathbf{P}}_j \quad). \end{aligned}$$

Here, Z_0 and Z_l are independent of each other whenever l is not contained in the set $\mathcal{I} := \{0, 1, \ldots, 52\} \cup \{n-53, n-52, \ldots, n-1\}$. In this case, $P[Z_0 = i \text{ and } Z_l = j] = P[Z_0 = i]P[Z_l = j]$ and the corresponding terms cancel in the covariance $Cov\left[\tilde{C}_i^{(n)}, \tilde{C}_j^{(n)}\right]$. Thus

$$\sigma_{ij} = \left(\sum_{l\in\mathcal{I}} P[Z_0 = i \text{ and } Z_l = j] - \sum_{l\in\mathcal{I}} P[Z_0 = i]P[Z_l = j]\right),$$

which depends only on a finite set of the random variables X_l^* and can be computed by summing up over all possible values of these variables. A *Mathematica* implementation of the covariance matrix, which exploits symmetries in the problem and employs combinatorical formulas to compute the number of occurrences of certain cases, needs about half an hour of CPU time on a DEC 3000 Alpha workstation to compute the covariance matrix Σ to

$$\Sigma \approx \begin{pmatrix} 3.24953 & -5.46111 & 2.21158 \\ -5.46111 & 9.40138 & -3.94027 \\ 2.21158 & -3.94027 & 1.72869 \end{pmatrix}.$$

A weak inverse of Σ computed by *Mathematica*'s built-in function `PseudoInverse` is given by

$$\Sigma^{-} \approx \begin{pmatrix} 2.908145042372 & 0.306521639158 & -3.21466668153 \\ 0.306521639158 & 0.084909927118 & -0.39143156628 \\ -3.214666681530 & -0.391431566276 & 3.60609824781 \end{pmatrix}.$$

6.4 A Weak Inverse for Multinomial Distributions

In the following Lemma we show that the diagonal matrix $diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$ is a weak inverse of the covariance matrix of the multinomial distribution $\mathcal{MN}(1, \mathbf{P})$. From this it follows that the asymptotic distribution of the chi-square statistic of K. Pearson can be derived from Corollary 3.3 as a special case.

Lemma 6.9 *Let $S = \{1, \ldots, m\}$ and let $\Sigma = (\sigma_{ij})_{(i,j)\in S^2}$ be the covariance matrix of a multinomial distribution $\mathcal{MN}(1, \mathbf{P})$. If $\overline{\Sigma} = (\overline{\sigma}_{ij})_{(i,j)\in S^2} = diag(\frac{1}{P_1}, \ldots, \frac{1}{P_m})$, then $\Sigma\overline{\Sigma}\Sigma = \Sigma$.*

Proof: Let $(\Sigma\overline{\Sigma}\Sigma)_{ij}$ denote the element in row i and column j of the matrix product $\Sigma\overline{\Sigma}\Sigma$,

$$(\Sigma\overline{\Sigma}\Sigma)_{ij} = \sum_{u=1}^{m}\sum_{v=1}^{m} \sigma_{iu}\overline{\sigma}_{uv}\sigma_{uj} = \sum_{u=1}^{m} \sigma_{iu}\frac{1}{P_u}\sigma_{vj}. \tag{6.3}$$

Distinguish the two cases that (i) $i = j$, and that (ii) $i \neq j$. As to (i), we have $\sigma_{xx} = P_x(1 - P_x)$ so that (6.3) becomes, owing to $\sum_{k\neq i} P_k = 1 - P_i$,

$$\begin{aligned} (\Sigma\overline{\Sigma}\Sigma)_{ii} &= \sum_{u\neq i}(-P_iP_u)\frac{1}{P_u}(-P_uP_i) + P_i(1-P_i)\frac{1}{P_i}P_i(1-P_i) = \\ &= P_i(1-P_i)\left[P_i + 1 - P_i\right] = P_i(1-P_i). \end{aligned}$$

As to (ii), $\sigma_{xy} = -P_xP_y$ and (6.3) becomes, owing to $\sum_{k\notin\{i,j\}} P_k = 1 - P_i - P_j$,

$$\begin{aligned} (\Sigma\overline{\Sigma}\Sigma)_{ij} &= \sum_{u\notin\{i,j\}}(-P_iP_u)\frac{1}{P_u}(-P_uP_j) + \\ &\quad + P_i(1-P_i)\frac{1}{P_i}(-P_iP_j) + (-P_iP_j)\frac{1}{P_j}P_j(1-P_j) = \\ &= P_iP_j\left[1 - P_i - P_j - (2 - P_i - P_j)\right] = -P_iP_j. \end{aligned}$$

■

Bibliography

[1] M.S. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc.*, Ser **B**(28):131–142, 1966.

[2] N. S. Altman. Bit–wise behavior of random number generators. *SIAM J. Sci. Stat. Comput.*, **9**(5):941–949, 1988.

[3] P. Billingsley. *Probability and Measure*. Wiley and Sons, New York, second edition, 1986.

[4] D.E. Boekee. *A generalization of the Fisher information measure*. Delft University Press, Delft, 1977.

[5] P. Bratley, B. Fox, and L.E. Schrage. *A Guide to Simulation*. Springer-Verlag, New York, 1983.

[6] M. Brunner. *Anwendungen der Risikomengentechnik zum Beweis von Konvergenzsätzen für Markovketten*. PhD thesis, Universität Salzburg, Österreich, 1988.

[7] N. Cressie and T. Read. Multinomial Goodness-of-fit Tests. *J. R. Statist. Soc. B*, **46**(3):440–464, 1984.

[8] I. Csiszár. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, **8**:85–108, 1963.

[9] J. Eichenauer-Herrmann. Statistical independence of a new class of inversive congruential pseudorandom numbers. *Math. Comp.*, **60**:375–384, 1993.

[10] J. Eichenauer-Herrmann, E. Herrmann, and S. Wegenkittl. A survey of quadratic and inversive congruential pseudorandom numbers. In H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, editors, *Monte Carlo and Quasi-Monte Carlo Methods 1996*, number 127 in Lecture Notes in Statistics, pages 66–97. Springer, New York, 1997.

[11] K. Entacher and S. Wegenkittl. The pLab picturebook: Load tests and ultimate load tests, part II: Subsequences. Report no. 2, pLab – reports, University of Salzburg, 1997. Available on the internet at `http://random.mat.sbg.ac.at/team/`.

[12] G. S. Fishman. Multiplicative congruential random number generators with modulus 2^{β}: An exhaustive analysis for $\beta = 32$ and a partial analysis for $\beta = 48$. *Mathematics of Computation*, **54**:331–344, 1990.

[13] G. S. Fishman and L. R. Moore. A statistical evaluation of multiplicative congruential random number generators with modulus $2^{31} - 1$. *Journal of the American Statistical Association*, **77**:129–136, 1982.

[14] G.S. Fishman and L.R. Moore. An exhaustive analysis of multiplicative congruential random number generators with modulus $2^{31} - 1$. *SIAM J. Sci. Statist. Comput.*, **7**:24–45, 1986. see also the Erratum, ibib. **7**(1986), p. 1058.

[15] M. Flahive and H. Niederreiter. On inversive congruential generators for pseudorandom numbers. In G.L. Mullen and P.J.-S. Shiue, editors, *Finite Fields, Coding Theory, and Advances in Communications and Computing*, pages 75–80. Dekker, New York, 1992.

[16] I. J. Good. The serial test for sampling numbers and other tests for randomness. *Proc. Cambridge Philosophical Society*, **49**:276–284, 1953.

[17] J. Hartung, B. Elpelt, and K. H. Klösener. *Statistik*. R. Oldenburg, Munich, 9th edition, 1993.

[18] P. Hellekalek. Inversive pseudorandom number generators: concepts, results, and links. In C. Alexopoulos, K. Kang, W.R. Lilegdon, and D. Goldsman, editors, *Proceedings of the 1995 Winter Simulation Conference*, pages 255–262. IEEE Press, Piscataway, N.J., 1995.

[19] P. Hellekalek. Good random number generators are (not so) easy to find. to appear in Mathematics and Computers in Simulation, 1998.

[20] A. B. Israel and T. Greville. *Generalized Inverses: Theory and Applications*. Wiley Interscience Publications. Wiley and Sons, New York, 1974.

[21] K. Jacobs. *Markov-Prozesse mit endlich vielen Zuständen*. Heidelberger Taschenbücher 98, Selecta Mathematica IV. Springer Verlag, 1972.

[22] P. Kafka, F. Österreicher, and I. Vincze. On powers of f-divergences defining a distance. *Studia Sci. Math. Hungar.*, 26:415–422, 1991.

[23] S. Karlin. *A second course in stochastic processes*. Academic Press, London, 1991.

[24] S. Kotz et al., editors. *Minimum discrimination information (MDI) estimation*, volume 5 of *Wiley Interscience Publication*, pages 527–529. John Wiley, 1985.

[25] S. Kullback. *Information Theory and Statistics*. John Wiley, New York, 1959.

[26] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[27] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer, New York, 1986.

[28] P. L'Ecuyer. Efficient and portable combined random number generators. *Comm. ACM*, **31**(6):742–774, 1988.

[29] P. L'Ecuyer, J.F. Cordeau, and R. Simard. Close-points spatial tests for random number generators. Submitted for publication, 1997.

[30] H. Leeb and S. Wegenkittl. Inversive and linear congruential pseudorandom number generators in empirical tests. *ACM Trans. Modeling and Computer Simulation*, **7**:272–286, 1997.

[31] F. Liese and I. Vajda. *Markov-Prozesse mit endlich vielen Zuständen*. Teubner-Texte zur Mathematik, Band *95*. Teubner, 1987.

[32] G. Marsaglia. A current view of random number generators. In L. Billard, editor, *Computer Science and Statistics: The Interface*, pages 3–10. Elsevier Science Publishers B.V., 1985.

[33] M. Matsumoto and Y. Kurita. Twisted GFSR Generators. *ACM Trans. Model. Comput. Simul.*, **2**(3):179–194, 1992.

[34] M. Matsumoto and Y. Kurita. Twisted GFSR generators II. *ACM Trans. Model. Comput. Simul.*, **4**:254–266, 1994.

[35] K. Matusita. Decision rules based on the distance for problems of fit. *Ann. Math. Statist.*, 26:631–640, 1955.

[36] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer, London, 1993.

[37] M. Z. Nashed, editor. *Generalized Inverses and Applications*, New York, 1976. Academic Press.

[38] J. Neyman. Contributions to the theory of the χ^2 test. Proceedings of the First Berkley Symposium on Mathematical Statistics and Probability, 1949.

[39] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods.* SIAM, Philadelphia, USA, 1992.

[40] H. Niederreiter. On a new class of pseudorandom numbers for simulation methods. *J. Comput. Appl. Math.*, **56**:159–167, 1994.

[41] H. Niederreiter. New developments in uniform pseudorandom number and vector generation. In H. Niederreiter and P. Jau-Shyong Shiue, editors, *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of *Lecture Notes in Statistics*. Springer, 1995.

[42] S. Orey. An ergodic theorem for Markov chains. *Z. W-theorie verw. Gebiete*, pages 174–176, 1962.

[43] F. Österreicher. The construction of least favourable distributions is traceable to a minimal perimeter problem. *Studia Sci. Math. Hungar.*, 17:341–351, 1982.

[44] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and its statistical applications, 1997. submitted.

[45] K. Pearson. Philos. Mag, Set. 5, **50**: 157–175, 1900.

[46] O. E. Percus and P. A. Whitlock. Theory and Application of Marsaglia's Monkey Test for Pseudorandom Number Generators. *ACM Transactions on Modeling and Computer Simulation*, **5** (2):87–100, 1995.

[47] C. R. Rao and S. K. Mitra. *Generalized Inverse of Matrices and its Applications*. Wiley Series in Probability and Mathematical Statistics. Wiley and Sons, 1971.

[48] T. Read and N. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer Series in Statistics. Springer Verlag, New York, 1988.

[49] V. Romanovsky. *Discrete Markov Chains*. Wolters-Noordhoff Publishing, Groningen, Netherlands, 1970.

[50] E. Seneta. *Non-negative matrices and Markov chains*. Springer Series in Statistics. Springer, New York, second edition, 1981.

[51] E. Stadlober and R. Kremer. Sampling from discrete and continuous distributions with c-Rand. In G. Pflug and U. Dieter, editors, *Simulation and Optimization*, volume 374 of *Lecture Notes in Economics and Math. Systems*, pages 154–162. Springer–Verlag, Berlin, 1992.

[52] E. Stadlober and F. Niederl. C-Rand: a package for generating nonuniform random variates. In *Compstat '94, Software Descriptions*, pages 63–64, 1994.

[53] I. Vattulainen, T. Ala-Nissila, and K. Kankaala. Physical models as tests of randomness. *Physical Review E*, 52(3):3205–3213, 1995.

[54] I. Vincze. On the concept and measure of information contained in an observation. In J. Gani and V.F. Rohatgi, editors, *Contributions to Probability*, pages 207–214. Academic Press, 1981.

[55] S. Wegenkittl. Empirical testing of pseudorandom number generators. Master's thesis, Universität Salzburg, Österreich, 1995. Available on the internet at `http://random.mat.sbg.ac.at/team/`.

[56] S. Wegenkittl. The pLab picturebook: Load tests and ultimate load tests, part I. Report no. 1, pLab – reports, University of Salzburg, 1997. Available on the internet at `http://random.mat.sbg.ac.at/team/`.

Curriculum Vitae

Persönliche Daten

Name	Mag. Stefan Wegenkittl
Geburtsdaten	8. Juli 1969, Salzburg
Eltern	Renate und Mag. Willibald Wegenkittl
Staatsbürgerschaft	Österreich
Familienstand	ledig
Zivildienst	1. 10. 1987–31. 5. 1988
Adresse	Institut für Mathematik, Universität Salzburg
	Hellbrunnerstraße 34, 5026 Salzburg, Österreich
	Telefon: +43 (0)662 8044 5329
	EMail: `Stefan.Wegenkittl@sbg.ac.at`

Ausbildung

1975–1987	Volksschule und Gymnasium in Salzburg
11. Juni 1987	Matura mit Auszeichnung
1988–1995	Studium der Mathematik (Schwerpunkt Statistik und Wahrscheinlichkeitstheorie) und der Computerwissenschaften in Salzburg
1991	erste Diplomprüfung in beiden Studien
14. Dezember 1995	Sponsion in Mathematik mit Auszeichung
seit 1996	Doktoratsstudium an der Universität Salzburg