# Properties of Random Fitness Landscapes and Their Influence on Evolutionary Dynamics

## - A Journey through the Hypercube -

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Stefan Nowak

aus Abu Dhabi

| | |
|---|---|
| Berichterstatter: | Prof. Dr. Joachim Krug,<br>Universität zu Köln |
| | Prof. Dr. Anton Bovier,<br>Rheinische Friedrich-Wilhelms-Universität Bonn |

Tag der mündlichen Prüfung: 23.11.2015

## Abstract

A fitness landscape is a theoretical concept in population genetics where a fitness value, which measures the reproductive success of an organism and is represented by a real number, is assigned to each genotype. Content of this thesis is the analytical and numerical study of stochastic models for fitness landscapes. The focus is on the landscape ruggedness and its influence on evolutionary dynamics. One proxy for the ruggedness is the number of local maxima, i.e., genotypes from which every mutation leads to lowered fitness. Another way to quantify ruggedness is the study of accessible paths, i.e., successions of mutations that increase the fitness monotonically. The question whether accessible paths exist can be interpreted as a kind of percolation problem. One model for evolutionary dynamics that will be used is the adaptive walk. In this model type, populations are treated as single entities that move through the space of genotypes according to certain probabilistic rules. They are closely related to both ruggedness measures as they follow accessible paths and terminate at local maxima. Furthermore, the individual-based Wright-Fisher model is used to study recombination of genotypes, interactions between individuals and the influence of the underlying fitness landscape on these mechanisms.

## Kurzzusammenfassung

Fitnesslandschaften sind ein theoretisches Konzept der Populationsgenetik bei dem jedem Genotypen eine reelle Zahl zugeordnet wird welche den reproduktiven Erfolg, die Fitness, des entsprechenden Organismus repräsentiert. Inhalt dieser Arbeit ist die analytische und numerische Untersuchung von stochastischen Modellen für Fitnesslandschaften. Das Hauptaugenmerk ist auf die Rauigkeit der Landschaften gerichtet und welche Auswirkungen diese auf evolutionäre Prozesse hat. Rauigkeit wird hauptsächlich auf zwei verschiedene Arten gemessen, nämlich durch die Anzahl lokaler Maxima, d.h. Genotypen von denen jede einzelne Mutation die Fitness verringert, und durch das Vorhandensein von zugänglichen Pfaden, d.h. Abfolgen von Mutationen bei denen die Fitness monoton erhöht wird. Letzteres kann auch als eine Art von Perkolationsproblem aufgefasst werden. Evolutionäre Prozesse werden zunächst durch sogenannte Adaptive Walks modelliert. In dieser Art von Modell wird eine Population als einzelnes Objekt betrachtet, das sich nach bestimmten stochastischen Regeln durch den Raum der Genotypen bewegt. Adaptive Walks folgen zugänglichen Pfaden und enden auf einem lokalen Maximum. Damit eng mit diesen Konzepten verbunden. Desweiteren wird das individuenbasierte Wright-Fisher Modell in dieser Arbeit verwendet um die Rekombination von Genotypen, Wechselwirkungen zwischen Individuen und den Einfluss der zugrunde liegenden Fitnesslandschaft auf diese Mechanismen zu untersuchen.

# Contents

# 1. Introduction

Treating interdisciplinary problems is common practice in statistical physics. It is often the case that models and methods that were developed in this field in order to describe, for instance, interacting particles can also be applied to systems of interacting animals, persons, cars, companies, and so on. The corresponding branches of science related to these examples are biology, sociology, traffic engineering and economics, respectively. In this thesis, evolutionary biology will be studied from a physicist's perspective. This means that evolutionary processes, as described below, will be represented by idealized mathematical models. Many concepts that arise from this treatment are closely related to systems of interacting spins, but there are also many similarities to computer science. This chapter explains these relationships between the different fields and introduces the concepts that will be used in this thesis.

## 1.1. Evolution in a Nutshell

Evolution is the change of lifeforms over generations. The basic ideas of the modern theory of evolution go back to the mid-19th century and Charles Darwin's famous book *On the Origin of Species* [1]. Changes of an organism manifest in modified observable traits, the *phenotype*. Today it is known that these changes can be actually ascribed to modifications of the organism's blueprint, the *genotype*. The genetic information is physically stored in Deoxyribonucleic acid (DNA) molecules that have a very complex structure. Simply speaking, the actual information corresponds to a certain arrangement of monomers, the *nucleobases*. Since there are four different types of nucleobases (cytosine, guanine, adenine, and thymine), an arrangement can also be represented by a sequence consisting of four possible letters (e.g., C, G, A, and T, corresponding to the initial letters of the nucleobases). When organisms reproduce, the genotype is inherited by the offspring. However, the offspring's DNA sequence might differ slightly from that of its parent if *mutations* occur. They are caused, for instance, by replication errors of the DNA molecule. Therefore, also the offspring's phenotype might be modified, which in turn can cause that it becomes worse or better adapted to its environment compared to the parent.

If a mutation in the offspring is beneficial, the organism will be more likely to survive and leaves on average more offspring with the mutated genotype to the next generation. This reproductive success is measured by the *fitness*. In the long run, individuals with high fitness will outnumber individuals with lower fitness, a process known as *natural selection*. Thus the structure of the whole population will change on timescales of several generations such that its overall fitness increases over time. This can lead to *fixation* of

a genotype, i.e., only one genotype remains in the population while individuals carrying different genotypes go extinct.

Apart from mutation and selection, also other mechanisms play a role in evolution. For instance, random fluctuations of the environment or the population affect the offspring production and hence the whole process. Another important example is the *recombination* of genotypes which provides a way to alter them independent of the occurrence of mutations.

## 1.2. Basic Concepts

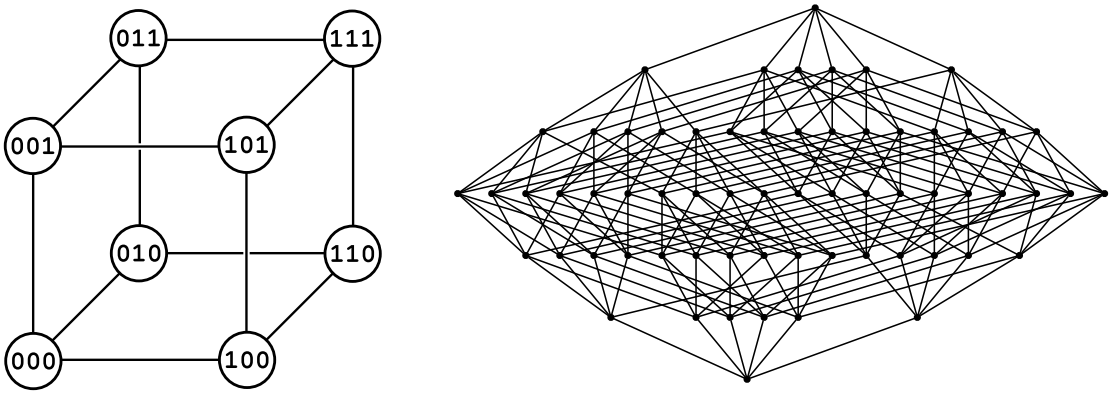
### 1.2.1. Space of Genotypes and the Hypercube

In the description above, genotypes correspond to DNA sequences consisting of letters from the set $\{\mathrm{C}, \mathrm{G}, \mathrm{A}, \mathrm{T}\}$. In general, sequences can also be made of letters from different sets that are denoted as *alphabets*. Different systems can be modeled with different alphabets, e.g., proteins have an alphabet of 20 letters corresponding to different amino acids. Throughout this thesis, genotypes will be represented by binary sequences of fixed length $L$ consisting of "letters" from the alphabet $A = \{0, 1\}$, i.e., the set of genotypes is given by $A^L$ and consists of $2^L$ sequences. The position of a letter in the sequence is called *locus*. A common biological interpretation is to denote the presence or absence of a certain mutation by one and zero, respectively. The loci correspond then to different possible mutations. In the same manner, genes that can occur as two different alleles or in two different states can be distinguished. A *mutation* at a certain locus corresponds to the change of a zero to a one or vice versa.

As a distance measure between two elements $\sigma$ and $\tau$ of $\{0, 1\}^L$, it is convenient to use the Hamming distance

$$d(\sigma, \tau) = L - \sum_{i=1}^{L} \delta_{\sigma_i, \tau_i} \,, \tag{1.1}$$

where $\delta_{i,j}$ is the Kronecker delta. This is nothing but the number of loci at which $\sigma$ and $\tau$ differ. Therefore, it is obvious that the Hamming distance is not restricted to the choice $A = \{0, 1\}$ as a reasonable metric. However, with this particular alphabet choice, an element $\sigma \in \{0, 1\}^L$ can be interpreted as a point of $\mathbb{R}^L$ that is located on a corner of the $L$-dimensional unit cube as shown in figure 1.1(a). For this reason, the undirected graph $\mathbb{H}_2^L$, where vertices correspond to sequences and edges correspond to two sequences at Hamming distance 1, is called hypercube graph or simply hypercube. Basic properties of this type of graph will be explained in section 2.1. Note that the graph topology does not depend on the particular choice of letters but only on their number. The generalization $\mathbb{H}_a^L$ to alphabets of arbitrary size $a = |A|$ is called Hamming graph.




**(a)** $L = 3$ **(b)** $L = 6$

**Figure 1.1.** Examples of hypercube graphs. Figure (a) is arranged in a way that demonstrates the cube structure. The graph in figure (b) can be arranged to a cube as well when embedded into 6-dimensional space, which is not obvious though.

### 1.2.2. Fitness Landscapes

Fitness is a measure for the reproductive success of an individual. It quantifies the individual's contribution to the gene pool of a population in succeeding generations. Therefore, fitness can be identified with the average number of offspring it leaves to the next generation. Independent of the interpretation, the fitness will be represented by a real number. Assuming that the genotype contains all information about the fitness of an individual, at least with respect to a given environment, there is a mapping $w\colon \mathbb{H}_2^L \to \mathbb{R}$ from the space of genotypes to the fitness which is called a *fitness landscape* [2–4]. Even though there exist some empirical fitness landscapes for small values of $L$ [3], the mechanism in which the genotype is translated into a fitness value is highly complex and far from being understood. For this reasons, models for fitness landscapes usually use random variables. Different variants that will be used in this thesis are defined in section 2.2.

#### Epistasis and Ruggedness

Landscapes on countrysides look very different in terms of ruggedness, e.g., they can be very plain, have gentle hills or be completely jagged. Ruggedness is usually associated with an increased number of mountain peaks. Hiking through such a landscapes often requires to go up- and downhill alternatingly. As the name suggest, fitness landscape can be thought of as a high-dimensional but discrete analogue by interpreting the fitness as a height profile. Mountain peaks correspond to local maxima, i.e., genotypes from which every single mutation leads to lower fitness. In the same way it is exhausting for a hiker to walk through a rugged landscape, it is difficult for a population to evolve in a rugged fitness landscape. Properties of fitness landscapes in terms of the local maxima and the existence of fitness-monotonic paths will be discussed in chapter 2 and 3, respectively.

Related to ruggedness is the concept of *epistasis* [5]. It means that the effect of a mutation at a certain locus is influenced by the state of other loci, the *genetic background*. In the absence of epistasis, a mutation at a certain locus will lower or increase the fitness by an amount that is independent of the state of other loci. Such a landscape is called *additive*, since the total fitness is simply the sum of the contributions from each locus. One distinguishes between different types of epistasis: *Magnitude epistasis* denotes the situation where only the strength of a mutational effect on the fitness is dependent on other loci, but not whether mutations are generally beneficial or deleterious. *Sign epistasis* [6], on the other hand, means that also the algebraic sign of a mutational effect varies, i.e., beneficial mutations can turn into deleterious ones (or vice versa) if mutations at different loci occur beforehand. High ruggedness is associated with a frequent occurrence of sign epistasis. *Reciprocal sign epistasis* [7] between two loci $i$ and $j$ means that they influence each other sign epistatically, i.e., $j$ has an sign epistatic effect on $i$ and vice versa.

### Landscapes Outside Biology

The concept of a fitness landscape can also be found in fields different from biology. A more general terminology is *value landscape*, a mapping $\mathcal{C} \to \mathbb{R}$ from the configuration space $\mathcal{C}$ of a system to the real numbers. The configuration space does not need to be a Hamming space, but it should include the notion of a neighborhood to justify the term "landscape".

This applies, for instance, to most physical systems where the energy plays the role of the (negative) fitness. In analogy to biological populations that are driven into states with high fitness, physical systems evolve into states with low energy. Metastable states correspond to local minima of the energy landscape. The analogy can be taken even further for systems of interacting spins [8]. If there are $L$ spins with two possible orientations $-1$ and $+1$, the configuration space $\mathcal{C} = \{-1, +1\}^L$ is isomorphic to the hypercube, where spin flips correspond to mutations.

Value landscapes also play an important role in computer science and optimization. On the *NK landscape*, a model for fitness landscapes that is also going to be used in this thesis, it is in general an NP-complete problem to determine whether the maximal fitness among all genotypes is below a given threshold [9]. For that reason, it is commonly used as a benchmark for optimization algorithms. Another famous example is the *traveling salesman problem*, where the task is to find the shortest possible route that visits each city on a given list of $n$ cities. The configuration space is then given by the set of permutations of cities while the (negative) fitness corresponds to the length of the route defined by a permutation [10, 11]. Similar to the situation on the NK landscape, it is an NP-hard problem to solve the traveling salesman problem [12] and hence there exist no efficient algorithm to find the global minimum of the corresponding landscape.

### 1.2.3. Evolutionary Dynamics

Evolutionary dynamics determines how the frequency of genotypes within a population changes over time. Models used to simulate evolutionary dynamics are often defined as stochastic processes. Since also the underlying fitness landscapes are modeled with random numbers, the outcome of a realization of the system is influenced by two different types of stochasticity. Certain questions can only be answered in terms of probabilities and averages.

In this thesis, two types of models for evolutionary dynamics are studied. They are both rather simple, yet they show interesting and non-trivial behavior. One type, a version of the *Wright-Fisher model* [13, 14], is individual based. Simply put, individuals produce offspring according to their fitness and mutations occur during this reproduction process with some probability, i.e., it corresponds basically to the scenario described in section 1.1. The dynamics will also be extended by the recombination of genotypes as well as with the competition between individuals. A detailed description of the model and its extensions will be given in chapter 5. The other type of dynamics, *adaptive walks*, is even simpler and arises in certain limits of the more general Wright-Fisher dynamics. Here one does not distinguish between individuals. Populations are rather treated as single objects that "walk" over the fitness landscape according to certain probabilistic rules. They will be studied in chapter 4.

Note that evolutionary dynamics plays also a role outside biology. For optimization problems, like the above mentioned traveling salesman problem, it is not feasible to search the whole configuration space. These and similar problems can be treated by so-called *genetic algorithms* [15] that mimic the behavior of a population evolving on the respective landscape. One can find particularly fit states with this method, which may suffice for practical applications, but in general one does not find the fittest state in large systems.

## 1.3. Structure of this Thesis

Chapter 2 begins with the recalling of basic properties of hypercube graphs and the definition of fitness landscapes models that are used in this thesis. Properties of these models will be discussed as well in this chapter with two different approaches. Firstly, a discrete analogue of Fourier analysis yields information about epistatic interactions on the landscape. Secondly, the study of local maxima serves as a direct proxy for landscape ruggedness. In particular, it will be studied how the scheme of epistatic interactions, which can be explicitly defined in the NK model, affects the number of maxima.

In chapter 3, *accessibility percolation* will be studied, a type of percolation problem that addresses paths to the global maximum such that the fitness is in ascending order along the paths. These accessible paths are favored to be taken by a population due to natural selection. In contrast to local maxima, which are local features of the landscape, accessible paths cross the whole landscape and are therefore global features.

Adaptive walks will be studied in chapter 4. Here it will be shown how the landscape properties influence the dynamics on it. For specific landscape models, a large class of

adaptive walks can be studied solely analytically. More sophisticated models require numerical simulations. Similar to the above-mentioned local maxima, the influence of epistatic interactions on the behavior of adaptive walks will be studied as well.

When the attention will be turned to the Wright-Fisher dynamics in chapter 5, results are almost entirely obtained by numerical simulations. The focus will be on two extensions of the basic dynamics, namely recombination and disruptive selection. In both cases, the underlying fitness landscape has crucial influence on the dynamics. In particular, the question of whether these mechanisms are advantageous for a population depends strongly on the landscape ruggedness.

A precise explanation of models and methods will be given in the beginning of each chapter. Apart from that, rather standard notation will be used throughout this thesis, but in order to avoid confusion and ambiguities it will also be explained in appendix A.1 on page 107.

# 2. Fitness Landscape Models and their Properties

## 2.1. Hypercubes

The hypercube graph, independently of a fitness landscape on top of it, has already interesting properties. Some of them will be recalled in the following.

- Since every vertex corresponds to a binary sequence of length $L$, there are $2^L$ vertices in total. Each of them has a degree of $L$ and hence there is a total of $L \cdot 2^{L-1}$ edges in the graph.
- The number of sequences at distance $d$ to a reference sequences $\sigma$ is given by $\binom{L}{d}$.
- The hypercube is a *bipartite graph*, i.e., the vertices can be divided into two disjoint sets such that there are no edges within a set. For instance, the sequences can be allocated according to whether they contain an odd or even number of ones.
- The hypercube is a *Hamiltonian graph*, i.e., there exist cycles in the graph that visit each vertex exactly once [16]. A famous example of such a cycle is the Gray code [16, 17].
- A path through the hypercube that contains $n$ vertices can be represented by the starting vertex and a string of length $n-1$ consisting of numbers in $\{1, 2, \ldots, L\}$. The number at the $i$-th position corresponds to the locus which is flipped from 0 to 1 or vice versa in the $i$-th step. For instance, the path

  $$(0,0,0) \to (0,1,0) \to (0,1,1) \to (1,1,1) \to (1,1,0)$$

  can be represented by the string `2313`. The path is self-avoiding if and only if there is no substring that contains all occurring numbers an even number of times.
- Between two vertices $\sigma$ and $\tau$ at distance $d = d(\sigma, \tau)$ there are $d!$ *shortest paths* that correspond to the number of possible successions in which the differing loci of $\sigma$ and $\tau$ can be flipped.
- The number of shortest paths between two sequences with $d(\sigma, \tau)$ that share exactly $k-1$ interior vertices (i.e., vertices different from $\sigma$ and $\tau$) is equal to the number $T(d,k)$ of *permutations with $k$ components* [18], a number that appeared in the mathematical literature before, independently of the hypercube context [19, 20].

Though no formula exist for $T(d,k)$ in simple closed form, it was proven in [18] that

$$d!\left(1-\mathcal{O}\left(\frac{1}{d}\right)\right) \leq T(d,1) \leq d!\,, \tag{2.1}$$

which means that most of the paths do not share any interior vertices for large $d$.

- The total *number of self-avoiding paths* (that are not required to be shortest) between two sequences $\sigma$ and $\tau$ is not known exactly. However, the question was mainly addressed to antipodal vertices, i.e., to the case $d(\sigma,\tau) = L$. The corresponding number $a_L$ was computed by simple enumeration of all paths up to $L = 5$ [21]. As $a_L$ grows very rapidly, this is not possible anymore for $L > 5$. In [22] it was found that $a_L$ grows double-exponentially, or more precisely, that

$$\lim_{L\to\infty} \frac{\log(\log a_L)}{L} = \log 2\,. \tag{2.2}$$

## 2.2. Random Models for Fitness Landscapes

### 2.2.1. House-of-Cards Model

The House-of-Cards (HoC) model [23, 24] is in some sense the simplest version of a random fitness landscape model. For each genotype $\sigma$, the fitness $w(\sigma)$ is a random number drawn independently from a continuous probability distribution. If the landscape is interpreted as an energy landscape, the HoC model is the analogue of Derrida's random energy model [25, 26].

### 2.2.2. Rough-Mount-Fuji Model

The Rough-Mount-Fuji (RMF) model [27, 28] is a simple extension to the HoC model in the version used here. It introduces a global gradient to the landscape, i.e., the fitness is given by

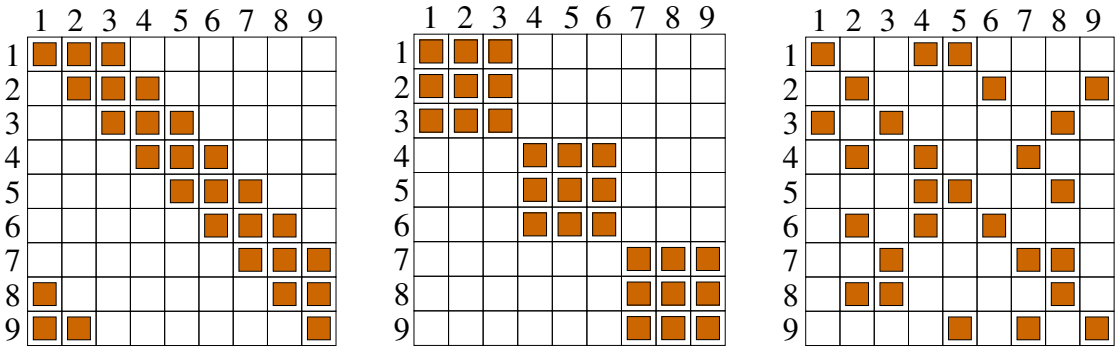
$$w(\sigma) = \eta(\sigma) + s\,d(\sigma,\tilde{\sigma})\,, \tag{2.3}$$

where $\eta$ is a HoC landscape, $\tilde{\sigma}$ is some reference sequence and $s$ is a constant. Given that $s > 0$, the fitness increases on average the further away $\sigma$ is from $\tilde{\sigma}$. The strength of the increase is controlled by the slope parameter $s$. The term $\eta(\sigma)$ will be referred to as the *random part* of the fitness, the second term as the *additive part*.

### 2.2.3. NK Model

Basic idea of this model is that total fitness associated with a genotype is made up of several contributions, where each contribution depends only on $1 \leq K \leq L$ loci of the genotype [29]. Fitness is given by

$$w(\sigma) = \sum_{i=1}^{L} \eta_i\big(\sigma_{b_{i,1}},\ldots,\sigma_{b_{i,K}}\big)\,, \tag{2.4}$$



**(a)** Adjacent **(b)** Block **(c)** Random

**Figure 2.1.** Illustration of the classic interaction patterns for $L = 9$ and $K = 3$. A filled square in row $i$ and column $j$ means that $V_i$ contains $j$.

where for each $i$ and each $\tau \in \{0,1\}^K$ the contribution $\eta_i(\tau)$ is a random number independently drawn from some continuous distribution. In other words, the $\eta_i$ are independent HoC landscapes of size $K$. The matrix $b_{i,j}$ determines the interactions between different loci, often referred to as the *genetic neighborhood*. The order of the arguments of $\eta_i$ does not affect landscape properties and hence the interaction scheme can be defined equivalently as sets

$$V_i = \{b_{i,1}, b_{i,2}, \ldots, b_{i,K}\}\,. \tag{2.5}$$

There are almost no constraints on these sets apart from their number being equal to $L$ and that $|V_i| = K$ as well as $i \in V_i$ for all $i$. The last condition ensures that each locus occurs in at least one set, but there is actually no mathematical or biological reason to assume that the number of contributions is equal to $L$ or that each contribution depends on the same number $K$ of loci. Note that in the literature the parameter $K$ is often defined in a way that the random functions $\eta_i$ depend on $K + 1$ rather than $K$ loci. Furthermore, the genotype length is often denoted by $N$ rather than $L$ (and hence the name "NK model" with regard to the parameters $N$ and $K$).

### Interaction Patterns

As mentioned above, there are lots of degrees of freedom concerning the interaction sets $V_i$. The most common choices in the literature, which will also be used in this thesis, are the following (see figure 2.1 for an illustration):

**Adjacent.** Tupels of $K$ adjacent loci interact with each other, i.e., the interaction sets are given by

$$V_i = \{i, i+1, \ldots, i+K-1\}\,, \tag{2.6}$$

where the elements have to be read modulo $L$.

**Block.** The sequence is divided into $L/K$ blocks, where each locus interacts with all other loci within the same block, but not with loci from other blocks. Interaction sets are given by

$$V_i = \left\{ K\left(\left\lceil \frac{i}{K} \right\rceil - 1\right) + 1, K\left(\left\lceil \frac{i}{K} \right\rceil - 1\right) + 2, \ldots, K\left(\left\lceil \frac{i}{K} \right\rceil - 1\right) + K \right\}, \quad (2.7)$$

where $\lceil x \rceil$ is the ceiling function. The sequence length $L$ has to be an integer multiple of $K$.

**Random.** The set $V_i$ contains $i$ as well as $K-1$ randomly chosen elements from $\{1, \ldots, i-1, i+1, \ldots, L\}$.

Depending on the property under consideration, the interaction scheme has more or less influence on the landscape. There are also certain characteristics, like the Fourier spectrum (see section 2.3), that are completely independent of the particular choice of the scheme. Other properties, e.g. the number of local maxima, are highly susceptible to this choice as will be shown later.

## 2.3. Fourier Decomposition

### 2.3.1. Expansion in Eigenfunctions of the Graph Laplacian

Any function $w\colon \mathbb{H}_2^L \to \mathbb{R}$ can be decomposed into eigenfunctions of the graph Laplacian $\Delta$ of the hypercube [30–32]. This transformation can be thought of as a discrete analogue of a Fourier transformation. The graph Laplacian reads $\Delta = \mathcal{A} - L\mathbb{1}_{2^L}$ where $\mathbb{1}_n$ is the $n \times n$ identity matrix and

$$\mathcal{A}_{\sigma,\tau} = \begin{cases} 1 & \text{if } d(\sigma,\tau) = 1\,, \\ 0 & \text{otherwise}\,, \end{cases} \quad (2.8)$$

is the adjacency matrix of the hypercube. Treating $\Delta$ as an operator, its effect on a function $w$ is given by

$$\Delta w(\sigma) = \sum_{\tau \in \mathbb{H}_2^L} \mathcal{A}_{\sigma,\tau} w(\tau) - L\, w(\sigma) = \sum_{\tau \in \mathcal{U}_\sigma} w(\tau) - L\, w(\sigma)\,, \quad (2.9)$$

where $\mathcal{U}_\sigma = \left\{\tau \in \mathbb{H}_2^L \mid d(\sigma,\tau) = 1\right\}$ is the set of neighbors of $\sigma$. The eigenfunctions of $\Delta$, also known as *Walsh functions*, are given by

$$\phi_I(\sigma) = 2^{-L/2} \cdot (-1)^{\sum_{i \in I} \sigma_i}\,, \quad (2.10)$$

where the set $I = \{i_1, \ldots, i_p\}$ of indices is a subset of $\{1, \ldots, L\}$. The corresponding eigenvalues $\lambda_p = -2p$ do only depend on $p$ and hence they have a rather large degeneracy of $\binom{L}{p}$. With the inner product defined by

$$\langle \phi, \psi \rangle = \sum_{\sigma \in \mathbb{H}_2^L} \phi(\sigma)\, \psi(\sigma)\,, \quad (2.11)$$

the Walsh functions form an orthonormal basis such that a function $w$ can be expressed as a linear combination

$$w(\sigma) = \sum_{p=0}^{L} \sum_{i_1<\cdots<i_p} a_{i_1,\ldots,i_p}\, \phi_{\{i_1,\ldots,i_p\}}(\sigma)\,. \tag{2.12}$$

This transformation is sometimes also called Walsh transformation [33, 34]. If the genotypes are represented by sequences from $\mathbf{s} \in \{-1,+1\}^L$ rather than $\sigma \in \{0,1\}^L$, which corresponds to a transformation $\mathbf{s} = 2\sigma - 1$, the Walsh functions take a particularly simple form proportional to products $s_{i_1} \cdots s_{i_p}$. Strings consisting of $-1$ and $+1$ can be naturally interpreted as configurations of one-dimensional spin systems. Therefore, the Fourier transformation (2.12) has the same form as the energy of a superposition of diluted $p$-spin glasses [25, 35]:

$$w(s) = w_0 + \sum_{i=1}^{L} h_i\, s_i + \sum_{p=2}^{L} \sum_{i_1<\cdots<i_p} J_{i_1\ldots i_p}\, s_{i_1} \cdots s_{i_p}\,, \tag{2.13}$$

where $h_i$ is a random "magnetic field" and $J_{i_1\ldots i_p}$ are random "coupling constants".

### 2.3.2. Amplitude Spectrum

The importance of the Fourier transformation is based on the exposure of interactions between loci. A coefficient $a_{i_1\ldots i_p}$ for $p > 1$ is only non-zero, if all loci $i_1, \ldots, i_p$ interact epistatically with each other. Its absolute value tells how strong the interaction is. On the other hand, the first-order coefficients $a_i$ tell how strong the non-epistatic influence of locus $i$ is, i.e., how strong the average effect of a mutation at locus $i$ is, independent of the state of other genes.

The *spectrum* $B_p$ is a way to quantify the influence of certain orders $p$ of interactions. Among other similar definitions, it can be defined as

$$B_p = \frac{\sum_{i_1<\cdots<i_p} \big\langle |a_{i_1,\ldots,i_p}|^2 \big\rangle}{2^L \operatorname{Cov}(w) + \sum_{q=1}^{L} \sum_{i_1<\cdots<i_q} \big\langle |a_{i_1,\ldots,i_p}|^2 \big\rangle}\,, \qquad p \in \{1,\ldots,L\}\,, \tag{2.14}$$

where $\langle \ldots \rangle$ means averaging over realizations of the landscape and

$$\operatorname{Cov}(w) = 4^{-L} \sum_{\sigma,\tau \in \mathbb{H}_2^L} \Big[ \big\langle w(\sigma)\, w(\tau) \big\rangle - \big\langle w(\sigma) \big\rangle \big\langle w(\tau) \big\rangle \Big] \tag{2.15}$$

is the mean covariance of the fitness. The amplitude spectrum $B_p$ is also related to the autocorrelation function $R_d$ via a linear transformation [30].

In the NK model, there are only epistatic interactions between loci $i_1, \ldots, i_p$ if they are all contained in at least one interaction set $V_j$. Since these sets contain only $K$ elements, it is obvious that $a_{i_1\ldots i_p} = 0$ for $p > K$ and hence the sum over $p$ in equations (2.12)

and (2.13) terminates at $p = K$. By definition, this implies that $B_p = 0$ for $p > K$. In fact, the spectrum can be computed explicitly [32] and is given by

$$B_p^{(\mathrm{NK})} = 2^{-K} \binom{K}{p}, \tag{2.16}$$

independent of the underlying interaction scheme of the NK landscapes. The spectrum of the HoC model is obtained from (2.16) by setting $K = L$. One finds for the RMF model that

$$B_p^{(\mathrm{RMF})} = \frac{N s^2 \, \delta_{p,1} + 4 \operatorname{Var}(\eta) \, 2^{-N} \binom{L}{p}}{N s^2 + 4 \operatorname{Var}(\eta)}, \tag{2.17}$$

where $\operatorname{Var}(\eta)$ is the variance of the random component of the landscape. This spectrum corresponds to a superposition of a HoC and an additive landscape.

### 2.3.3. The Rank

The *rank* $R$ was introduced in [36] to study interaction patterns of NK landscapes. Among other equivalent definitions, it can be defined as the number of non-zero coefficients in the Fourier expansion (2.12). Therefore, it is not applicable to HoC or RMF landscapes, since the random term of the fitness causes that all coefficients are non-zero which leads to $R = 2^L$.

For NK landscapes, however, it will turn out to be very useful for the quantification of interaction schemes since the rank, unlike the spectrum, is strongly influenced by them. For the actual calculation of the rank, it is convenient to use the equivalent definition

$$R = |\mathcal{V}| = \left| \bigcup_{i=1}^{L} \mathcal{P}(V_i) \right|, \tag{2.18}$$

where $\mathcal{P}(S)$ and $|S|$ denote power set and counting measure, respectively, of a set $S$. The set $\mathcal{V} = \bigcup_i \mathcal{P}(V_i)$ contains all sets of indices of non-zero coefficients in the Fourier expansion.

The maximal rank of a classic NK landscapes with fixed $L$ and $K$ is reached when the overlap between interaction sets is as small as possible. Each interaction set $V_i$ can contribute at most $\binom{K}{m}$ subsets of size $m$ to the union $\mathcal{V}$. Since the empty set and the $L$ unit sets are always contained in $\mathcal{V}$, an upper limit for the rank is given by

$$R_{\max} = 1 + L + L \sum_{m=2}^{K} \binom{K}{m} = 1 + L \left(2^K - K\right). \tag{2.19}$$

This is a sharp bound and hence one can construct interaction schemes that reach this rank. However, not all values of $L$ and $K$ allow for such patterns with $R = R_{\max}$. It can be shown that such a scheme exists if an $(L, K)$-packing design, an object known in combinatorial design theory, exists [36]. The actual rank for the different standard interaction patterns was computed in [37] and will be presented in the following.

### Rank of Block Interactions

It is particularly easy to calculate the rank for blockwise interactions, because two contributions $\mathcal{P}(V_i)$ and $\mathcal{P}(V_j)$ to equation (2.18) are always either equal or disjoint apart from the empty set. The number of disjoint contributions is equal to the number $L/K$ of blocks and each contribution contains $2^K$ elements. Taking into account that the empty set is only counted once, the rank is given by

$$R_{\text{block}} = \frac{L}{K}\big(2^K - 1\big) + 1\,. \tag{2.20}$$

### Rank of Adjacent Interactions

In the adjacent case, the interaction set $V_i$ contains the integers from $i$ to $i+K-1$, but all elements are taken modulo $L$. Therefore, one can interpret the elements as particle positions on a one-dimensional lattice of length $L$ with periodic boundary conditions. By definition, the particles are strung together and form a cluster of size $K$. Therefore, a set $S$ is contained in $\mathcal{V}$ if and only if all particle positions in $S$ can be found in an interval of size $K$, or equivalently, if there is a gap of size $L-K$ or larger. The rank is nothing but the number of sets contained in $\mathcal{V}$ and thus it is equal to the number of ways one can put an arbitrary amount of indistinguishable particles on a periodic lattice such that a gap of size $L-K$ emerges.

For $K < (L+1)/2$, there can be only one gap of the required size such that it is straightforward to enumerate all valid particle configurations. One has exactly $L$ possibilities to choose the first occupied site after the gap. Then the $K-1$ subsequent sites can be either occupied or empty, giving a factor of $2^{K-1}$ to each position of the first occupied site. Finally, the configuration without any particles has to be included as well. This leads to

$$R_{\text{adj}} = 1 + L\,2^{K-1}\,, \qquad \text{if} \quad K < (L+1)/2\,. \tag{2.21}$$

If $K$ is too large, there may be several gaps. In this case, the factor $2^{K-1}$ also includes configurations with another gap of the required size. As a consequence, these configurations are counted more than once by the enumeration procedure described above and equation (2.21) overestimates the actual rank.

### Rank of Random Interactions

By definition, the actual rank for random interaction is a stochastic quantity. Its mean value will be calculated in the following. Let $S$ be an arbitrary subset of $\{1,\dots,L\}$ with $|S| = m$ and $p_m$ the probability that $S$ is contained in $\mathcal{V}$. Unit sets as well as the empty set are always contained and hence $p_0 = p_1 = 1$. For $m > 1$ one finds

$$\begin{aligned} p_m &= \mathbb{P}[S \in \mathcal{V}] = \mathbb{P}[\exists\, i\colon S \subset V_i] \\ &= \mathbb{P}[\exists\, i \in S\colon S \subset V_i] + \mathbb{P}[\exists\, i \notin S\colon S \subset V_i] \\ &= \big[1 - (1-q_m)^m\big] + \big[1 - (1-q'_m)^{L-m}\big]\,, \end{aligned} \tag{2.22}$$

where

$$q_m = \frac{\binom{K-1}{m-1}}{\binom{L-1}{m-1}} = \frac{(K-1)!\,(L-m)!}{(L-1)!\,(K-m)!}$$

and

$$q'_m = \frac{\binom{K-1}{m}}{\binom{L-1}{m}} = q_m\,\frac{K-m}{L-m}$$

are the probabilities that $S \subset V_i$, conditioned on $i \in S$ and $i \notin S$, respectively. When $K$ is sufficiently smaller than $L$, the probabilities $q_m$ and $q'_m$ are very small and one can use the approximation $(1-x)^n \approx 1 - nx$ in equation (2.22) that yields

$$p_m \approx m\,q_m + (L-m)\,q'_m = K\,q_m = \frac{K!\,(L-m)!}{(L-1)!\,(K-m)!}\,. \tag{2.23}$$

Finally, the mean rank is obtained by summing over all possible sets which reads

$$\begin{aligned}\mathbb{E}[R_{\mathrm{rnd}}] &= \sum_{S\subset\{1,\ldots,L\}} \mathbb{P}[S \in \mathcal{V}] = 1 + L + \sum_{m=2}^{K} \binom{L}{K} p_m \\ &\approx 1 + L + \sum_{m=2}^{K} \binom{L}{K} \frac{K!\,(L-m)!}{(L-1)!\,(K-m)!} \\ &= 1 + L(2^K - K) = R_{\max}\,. \end{aligned} \tag{2.24}$$

Since the average rank is close to the upper limit, it is safe to assume that the fluctuations around $\mathbb{E}[R_{\mathrm{rnd}}]$ are very small.

## 2.4. Local Maxima

Local maxima of a fitness landscape are those sequences that only have neighbors with lower fitness. They play an important role for evolutionary dynamics, because individuals whose genotype is a local maximum cannot improve their fitness by a single mutation which makes it hard to escape from them. The total number or density of local maxima is an important measure for the ruggedness of the landscape. Furthermore, it is also important to know how the maxima are distributed over the landscape as their positions are correlated. For example, it is by definition not possible that two neighboring genotypes are both maxima, but apart from that they tend to form clusters in most models.

### 2.4.1. HoC Model

In this model, the independence of fitness values facilitates the study of local maxima by a large amount. A sequence $\sigma$ being fitter than all of its $L$ neighbors is equivalent

to $w(\sigma)$ being the largest of $L+1$ i.i.d. random variables. Since each random variable has the same chance to be the largest, the probability $P_{\max}$ that a genotype is a local maximum is independent of the underlying fitness distribution and given by

$$P_{\max} = \frac{1}{L+1}\,. \tag{2.25}$$

Multiplying $P_{\max}$ with the total number of genotypes gives the expected number of local maxima

$$\mathbb{E}[N_{\max}] = \frac{2^L}{L+1}\,. \tag{2.26}$$

The fitness $h$ of a local maximum can be obtained due to the fact that it is the largest of $L+1$ random variables as well. Its cumulative distribution function (CDF) reads

$$F_{\max}(x) = F(x)^{L+1}\,, \tag{2.27}$$

where $F$ is the overall fitness distribution function. In contrast to $N_{\max}$, the average height $\mathbb{E}[h]$ of local maxima and even the scaling with $L$ depends on the distribution. One can, however, always scale the fitness values to the uniform distribution which yields

$$\mathbb{E}[F(h)] = 1 - \frac{1}{L+2}\,. \tag{2.28}$$

Moreover, one can compute the probability $P_{\max,2}$ that two sequences $\sigma$ and $\tau$ with $d(\sigma,\tau) = 2$ are both local maxima. They do not have an independent chance to be local maxima as their neighborhoods share two common vertices. The number of genotypes involved in this situation is $2L$ and hence both $\sigma$ and $\tau$ have a probability of $1/2L$ to be the largest of them. Given that, for instance, $\sigma$ has the largest fitness, the probability that $\tau$ is larger than its neighbors is still $1/(L+1)$ and hence

$$P_{\max,2} = 2 \cdot \frac{1}{2L} \cdot \frac{1}{L+1} = \frac{1}{L\,(L+1)} > P_{\max}^2\,, \tag{2.29}$$

where the factor 2 is due to the interchangeable roles of $\sigma$ and $\tau$. The result implies that there is a weak enhancement in probability to find local maxima at distance 2 to other maxima, even though fitness values are uncorrelated. For $d(\sigma,\tau) > 2$, however, $\sigma$ and $\tau$ have an independent chance to be local maxima since they do not share any common neighbors.

### 2.4.2. RMF Model

In contrast to the other models, the RMF model is not isotropic in the sense that the probability $P_{\max}$ for a sequence $\sigma$ depends on its distance $d(\sigma,\tilde{\sigma})$ to the reference sequence. Suppose $d(\sigma,\tilde{\sigma}) = d$, then the fitness $w(\sigma)$ has according to equation (2.3) the cumulative distribution function $F(x - s\,d)$, where $F$ is the CDF of the random part of the fitness. Furthermore, $\sigma$ has $d$ neighbors that are located closer to $\tilde{\sigma}$ (downhill) and

$L-d$ neighbors that are farther away (uphill). Obviously, all down- and uphill neighbor must have smaller fitness than $w(\sigma)$ if $\sigma$ is a local maxima. The probability for this is given by $F[w(\sigma)-s\,(d-1)]^d \cdot F[w(\sigma)-s\,(d+1)]^{L-d}$. Integrating over $\sigma$'s possible fitness values yields

$$P_{\max}(d) = \int_{-\infty}^{\infty} f(x)\, F(x+s)^d\, F(x-s)^{L-d}\,\mathrm{d}x\,. \tag{2.30}$$

Since $F(x+s) > F(x-s)$, one can see immediately that $P_{\max}(d)$ increases with $d$. The corresponding probability for a randomly chosen genotype is accordingly given by

$$P_{\max} = 2^{-L}\sum_{d=0}^{L}\binom{L}{d}P_{\max}(d) = 2^{-L}\int_{-\infty}^{\infty} f(x)\,[F(x+s)+F(x-s)]^L\,\mathrm{d}x\,. \tag{2.31}$$

Equations (2.30) and (2.31) cannot be expressed in simple closed form for arbitrary distributions of the random part, but it was shown in [38] that the leading order behavior of $P_{\max}$ is given by

$$P_{\max} = \frac{1}{L+1} - \frac{s^2\,L\,(L-1)}{2}\int_{-\infty}^{\infty} f(x)^3 F(x)^{L-2}\,\mathrm{d}x + \mathcal{O}\big(s^4\big)\,. \tag{2.32}$$

Furthermore, a couple of special cases were presented in [38] where $P_{\max}(d)$ or $P_{\max}$ can be computed asymptotically exact. An interesting example is the Weibull distribution

$$F(x) = \big[1-e^{-x^\nu}\big]\theta(x)\,, \tag{2.33}$$

where $\theta(x)$ is the Heaviside function. It was shown that the asymptotic behavior is given by

$$P_{\max} \sim \begin{cases} 1/L & \text{for } \nu < 1\,, \\ \frac{e^s\left[\left(1-e^{-2s}\right)^{L+1}-\left(1-e^{-s}\right)^{L+1}\right]}{2^L\,(L+1)} + \frac{1-\left(1-e^{-s}\cosh s\right)}{(L+1)\,\cosh s} & \text{for } \nu = 1\,, \\ \frac{1}{L}\exp\big[-\nu\, s\,(\log L)^{1-1/\nu}\big] & \text{for } \nu > 1\,. \end{cases} \tag{2.34}$$

The result for $\nu < 1$ corresponds to the HoC model, which means that the fitness values are dominated by the random part if its distribution decays more slowly than exponentially and $L$ is large. On the contrary, for $\nu \geq 1$ the slope $s$ has also for $L \to \infty$ a notable effect.

### 2.4.3. NK Model

Obtaining results on the number of maxima in the NK model is a challenging task. Obviously, for a sequence to be a local maximum, each mutation must result in lower fitness which makes things complicated due to the change of several contributions at once. Nevertheless, one can write down the corresponding probability and the expected height of a local maximum, at least formally.

Suppose an NK landscape with interaction sets $V_i$, a sequence $\sigma$ with fitness

$$w(\sigma) = \sum_{i=1}^{L} x_i \,,$$

where $x_i = \eta_i\big(\sigma_{b_{i,1}}, \ldots, \sigma_{b_{i,K}}\big)$ are the contributions to the total fitness and let $f$ be the probability density function (PDF) of the contributions, i.e., the distribution of the total fitness is given by the $L$-fold convolution of $f$. When a mutation at the $i$-th locus occurs, the set of indices of affected contributions is given by

$$U_i = \big\{ j \in \{1, \ldots, L\} \mid i \in V_j \big\} \,, \tag{2.35}$$

i.e., all contributions $x_j$ with $j \in U_i$ will be altered to a new value $x'_j$. If $\sigma$ is a local maximum, the sum over the new values has to be smaller than the sum over the old values. Assuming that the $x_i$ are fixed, the probability for this can be written as

$$\mathbb{P}\left[ \sum_{j \in U_i} x'_j < \sum_{j \in U_i} x_j \right] = \tilde{F}_{|U_i|}\Big( \sum\nolimits_{j \in U_i} x_j \Big) \,, \tag{2.36}$$

where $\tilde{F}_n$ is the cumulative distribution function corresponding to the probability density defined by the convolution

$$\tilde{f}_1(x) = f(x) \,, \quad \tilde{f}_{n+1}(x) = \int_{-\infty}^{\infty} \mathrm{d}z \, f(z) \, \tilde{f}_n(x - z) \,, \tag{2.37}$$

i.e., $\tilde{f}_n$ is the PDF of $\sum_{j \in U_i} x'_j$ for $|U_i| = n$. Note that mutations at different loci yield different sets of new contributions and hence, as long as the vector $\mathbf{x} = (x_1, \ldots, x_L)$ of old contributions is fixed, the probabilities in equation (2.36) are independent for different $i$. Therefore, the actual probability that $\sigma$ is a local maximum is obtained by taking the product over all $i$ and integrating over all values of $\mathbf{x}$. This reads

$$P_{\max} = \int_{\mathbb{R}^L} \mathrm{d}^L\mathbf{x} \prod_{i=1}^{L} \Big[ f(x_i) \, \tilde{F}_{|U_i|}\Big( \sum\nolimits_{j \in U_i} x_j \Big) \Big] \,. \tag{2.38}$$

Note that a similar expression, which was restricted to Gaussian random numbers, was derived in [36].

The integrand of equation (2.38) is the joint probability of being a local maximum and probability density of the contributions. Using the definition of conditional probabilities, the expected fitness $h$ of a sequence, given that it is a maximum, reads

$$\mathbb{E}[h] = \frac{1}{P_{\max}} \int_{\mathbb{R}^L} \mathrm{d}^L\mathbf{x} \left( \sum_{k=1}^{L} x_k \right) \prod_{i=1}^{L} \Big[ f(x_i) \, \tilde{F}_{|U_i|}\Big( \sum\nolimits_{j \in U_i} x_j \Big) \Big] \,. \tag{2.39}$$

In most cases, however, neither equation (2.38) nor (2.39) can be expressed in simple closed form. It is even worse for random interactions since $P_{\max}$ and $\mathbb{E}[h]$ depend on

the precise realization of the interaction scheme. If the interaction scheme is chosen randomly, $P_{\max}$ will be a random variable as well. One might also want to perform the average over realizations of the $U_i$, but this is obviously a difficult task. In the following $P_{\max}$ denotes nevertheless this average rather than the actual variable if not stated otherwise, i.e., it is defined as the probability that a randomly chosen genotype is a local maximum in a landscape with a randomly chosen interaction pattern. The same applies to $\mathbb{E}[h]$.

### Block Interactions

In the block model, in contrast to other interaction schemes, it is straightforward to compute the mean number of maxima [39, 40] and even the probability to have two maxima at certain distances [37]. A sequence $\sigma$ is a local maximum if and only if each of the sub-landscapes induced by the block structure has a maximum at the corresponding sub-sequence. There are $L/K$ blocks, each block's sub-landscape is a HoC landscape of size $K$, and hence

$$P_{\max} = \left(\frac{1}{K+1}\right)^{\frac{L}{K}} . \tag{2.40}$$

Given that $\sigma$ is a maximum, the probability that a second sequence $\tau$ at distance $d(\sigma,\tau) = 2$ is also a maximum depends on two things: Firstly, both loci in which $\sigma$ and $\tau$ differ have to be within the same block, otherwise the condition that each block has to be a maximum by itself cannot be fulfilled. If $\tau$ is randomly chosen among the second-nearest neighbors of $\sigma$, the probability for this is $(K-1)/(L-1)$. Secondly, the two alterations of the sequence have to lead to another maximum of the sub-landscape in the corresponding block. This happens with probability $1/K$ according to equation (2.29). Combining both considerations, the probability that two randomly chosen sequences at distance 2 are both maxima is given by

$$P_{\max,2} = \frac{K-1}{K\,(L-1)} P_{\max} , \tag{2.41}$$

which is very large compared to $P_{\max}^2$.

In principle, one can extend this method to sequences at arbitrary distance $d$. It is required that there are either no or at least two differing loci in each block. In case $d = 3$, all differing loci have to be in the same block and hence the calculation is completely analogous to the previous one. It yields

$$P_{\max,3} = \frac{(K-1)\,(K-2)}{(K+1)\,(L-1)\,(L-2)} P_{\max} = \frac{K\,(K-2)}{(K+1)\,(L-2)} P_{\max,2} . \tag{2.42}$$

For $d > 3$, a laborious case analysis is required since the differing loci can be distributed over the blocks in various ways.

| Distribution | $K$ | $\lambda_K$ | $\mathbb{E}[h]/L$ | Ref. |
|---|---|---|---|---|
| Gamma$(2,1)$ | 2 | 0.56457... | 2.88039... | [43] |
| Exp(1) | 2 | 0.56268... | 1.61651... | [43] |
| Exp(1) | 3 | 0.61140... | 1.86367... | [43] |
| Negative Exp(1) | 2 | 0.57695... | −0.48097... | [41] |
| Gamma$(1/2,1)$ | 2 | 0.56062... | 0.92242... | [37], this thesis |

**Table 1.** Exactly known values for the asymptotics of the mean number and height of local maxima in the NK model with adjacent interactions. The constant $\lambda_K$ is defined in equation (2.43).

### General Phenomenology

Many quantities on the NK landscape with adjacent or random interactions behave qualitatively like the model with block interactions. It is known for adjacent interactions that the number of maxima grows exponentially with sequence length $L$, or more precisely that

$$\lim_{L\to\infty} \frac{\log P_{\max}}{L} = \log \lambda_K \,, \tag{2.43}$$

with $1/2 < \lambda_K < 1$ being a constant [41] depending in general on the underlying distribution of fitness contributions. Obviously, the block model shows the same asymptotic behavior and one has $\lambda_K = (K+1)^{-1/K}$ according to equation (2.40), independent of the underlying fitness distribution. Apart from that, a few values of $\lambda_K$ for adjacent interactions and small values of $K = 2$ or $K = 3$ are known exactly. For the mean height of a local maximum it was found that it decreases proportional to $\sqrt{\log(K)/K}$ for large $K$ in case of Gaussian fitness contributions [42], which was conjectured to be also true for other distributions with finite variance [41]. However, exact results are as rare as for $P_{\max}$. Cases where $\lambda_K$ and the asymptotics of $\mathbb{E}[h]$ are known exactly are listed in table 1. One example, where they can be computed explicitly from equations (2.38) and (2.39), will be shown in the next section.

In the following, $P_{\max}$ and $\mathbb{E}[h]$ will be studied mostly with the numerical integration of (2.38) and (2.39). The algorithm used for the integration can be briefly explained as Monte-Carlo integration with importance sampling. A detailed description can be found in appendix A.2.2. It should be noted, however, that it works much better with the formula for $P_{\max}$ than for $\mathbb{E}[h]$, i.e., the algorithm needs substantially more sampling points for the computation of $\mathbb{E}[h]$ than for $P_{\max}$ in order to reach the same precision. This is enhanced by the fact that changes of $P_{\max}$ due to changes of the landscape parameters are usually much larger than the fluctuations of the algorithm, which is not always the case for $\mathbb{E}[h]$. For that reason, $\mathbb{E}[h]$ will be shown with error bars while the error of $P_{\max}$ is always much smaller than symbol sizes in all figures shown here. Of course, the results for $\mathbb{E}[h]$ can be improved with more computation time and/or a better algorithm, but this is left for future work.

The numerical integration suggests that equation (2.43) is also valid for random interactions, as shown in figure 2.2(a). Furthermore, both figures 2.2(a) and (b)

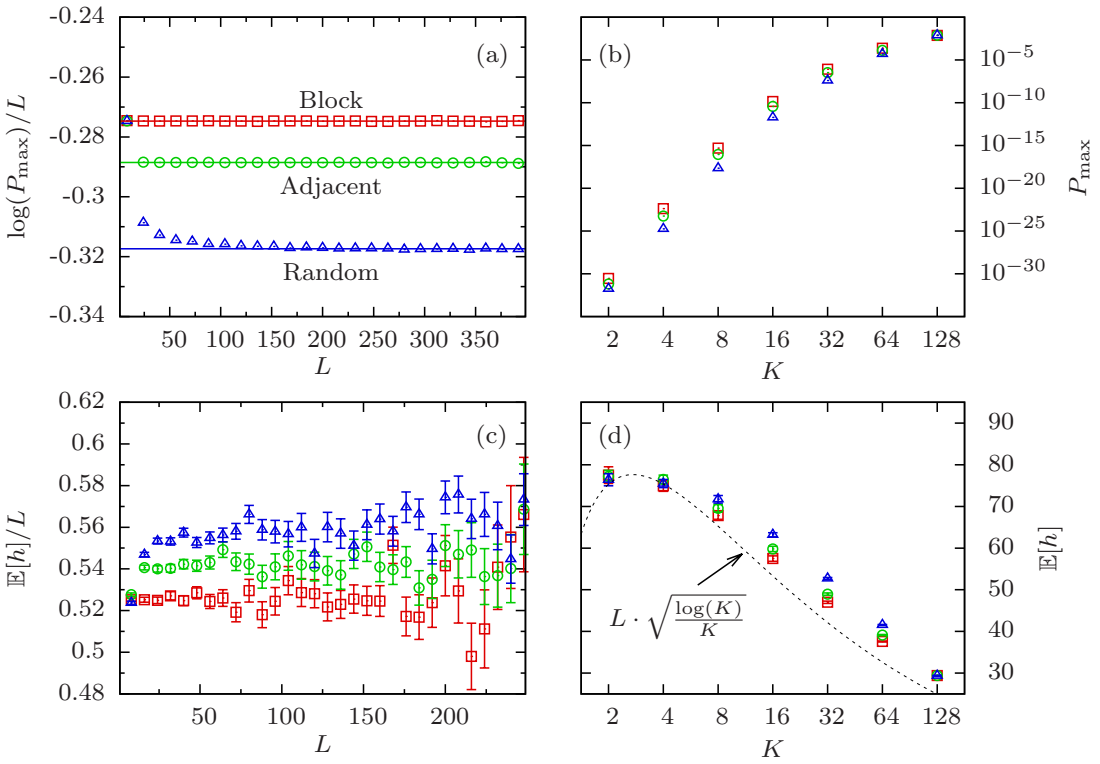



**Figure 2.2.** Study of $P_{\max}$ and $\mathbb{E}[h]$ on the NK landscape for fixed $K = 8$ (panel (a) and (c)) and fixed $L = 128$ (panel (b) and (d)), respectively. Symbols are defined in panel (a) and correspond to the numerical evaluation of equation (2.38) using standard normal distributed fitness contributions. Random interactions are averaged over 100 realizations.

show that $P_{\max}$ is smallest for random interactions while it is largest for block interactions. Adjacent interactions are roughly halfway in between. Because $P_{\max}$ decays exponentially with different base $\lambda_K$, the relative difference between interaction patterns grows exponentially with $L$. For that matter, the choice of interactions has a huge impact on the number of maxima, despite contrary statements in the literature [42]. However, $P_{\max}$ is still much more influenced by the parameter $K$ than by the type of interactions. Numerically obtained values for the mean height $\mathbb{E}[h]$ are shown in figure 2.2(c) and (d). Although errors due to the integration algorithm are quite large, one can observe that the order of heights with respect to different interaction schemes is the opposite of the order for $P_{\max}$. This means that random and adjacent interactions have less maxima than the block model, but their maxima have larger average fitness.

The special feature of random interactions was already mentioned briefly: One can treat $P_{\max}$ and $\mathbb{E}[h]$ either as the average over realizations of interaction schemes or as a random variable that takes different values depending on the realization. If the latter is assumed, one may ask how large the fluctuations are. The answer is partially given in figure 2.3(a) in terms of the standard deviation of $P_{\max}$. As one can see, the

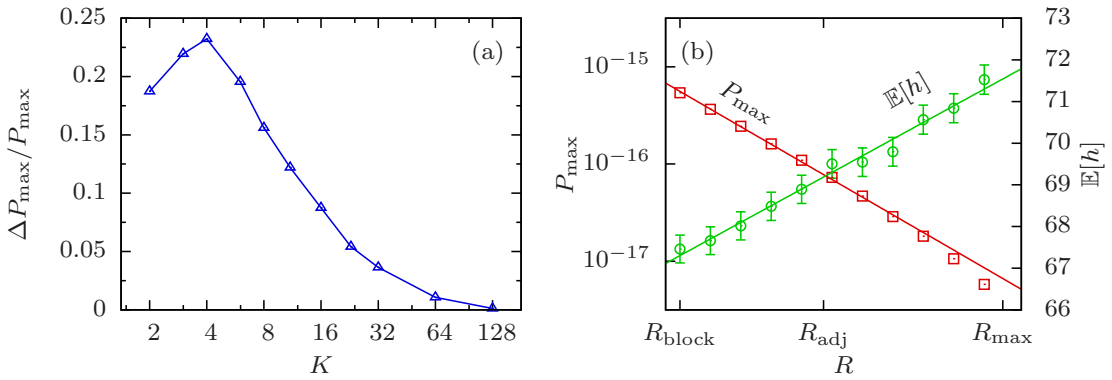



**Figure 2.3.** (a) Standard deviation $\Delta P_{\max}$ of $P_{\max}$ with random interactions for fixed $L = 128$ and varying $K$. Lines are for visual guidance. (b) Dependence of $P_{\max}$ and $\mathbb{E}[h]$ on the rank $R$ for fixed $L = 128$ and $K = 8$. Interaction schemes with certain rank are created with help of the algorithm described in appendix A.2.1. Each point corresponds to the average over 100 realizations. Lines are linear regressions.

fluctuations decline when $K$ is of the same order as $L$, but they are quite significant when $K$ is small. Interestingly, they are maximal for a value slightly above $K = 2$, resulting in non-monotonic behavior. Note, however, that the fluctuations are still very small compared to the difference between the averaged $P_{\max}$ for random interactions and $P_{\max}$ for adjacent interactions. It would be nice to study the fluctuations of $\mathbb{E}[h]$ as well, but they are hard to obtain since fluctuations caused by the integration algorithm are much larger than fluctuations due to the interaction scheme in that case.

Obviously, there are significant differences between the three classic interaction schemes in terms of $P_{\max}$ and $\mathbb{E}[h]$. What about interaction schemes that are not classic? At this point it turns out that the rank $R$ is a powerful tool for the quantification of arbitrary schemes [36, 37]. The dependence of $P_{\max}$ and $\mathbb{E}[h]$ on $R$ for fixed $L$ and $K$ is shown in figure 2.3(b). Although two schemes with the same rank might still have different properties, there is a rather clear correlation between these quantities and $R$. The (averaged) probability $P_{\max}$ decays roughly exponentially with $R$ while $\mathbb{E}[h]$ increases linearly. When $R$ is increased, the landscape properties seem to change smoothly from block model to the random model.

### Exactly Solvable Example of Adjacent Interactions

In the following, the example of a *Gamma distribution* with shape parameter $1/2$, $K = 2$, and adjacent interactions will be presented. The result for $P_{\max}$ was previously obtained in [37]. A gamma distribution with shape parameter $1/2$ has a PDF given by

$$f(x) = \begin{cases} \frac{\exp(-x)}{\sqrt{\pi x}} & x > 0\,, \\ 0 & \text{otherwise.} \end{cases} \tag{2.44}$$

Its self-convolution corresponds to an exponential distribution, i.e.,

$$\tilde{F}_2(x) = 1 - \exp(x)\,. \tag{2.45}$$

By introducing the auxiliary function

$$Z(\vartheta) = \frac{1}{\sqrt{\pi^L}} \int_{\mathbb{R}_+^L} \mathrm{d}^L\mathbf{x} \prod_{i=1}^{L} \left[ \frac{\exp(-\vartheta x_i)}{\sqrt{x_i}} \left(1 - e^{-x_i - x_{i+1}}\right) \right], \tag{2.46}$$

one can write $P_{\max}$ and $\mathbb{E}[h]$ according to equations (2.38) and (2.39) as

$$P_{\max} = Z(1) \qquad \text{and} \qquad \mathbb{E}[h] = -\frac{1}{Z(1)} \frac{\mathrm{d}Z(\vartheta)}{\mathrm{d}\vartheta}\bigg|_{\vartheta=1}, \tag{2.47}$$

respectively. To evaluate $Z(\vartheta)$, the second factor in brackets can be expanded as

$$\prod_{i=1}^{L} \left(1 - e^{-x_i - x_{i+1}}\right) = \sum_{\mathbf{s}\in\{0,1\}^L} \left( \prod_{i=1}^{L} (-1)^{s_i} e^{-x_i(s_{i-1}+s_i)} \right) \tag{2.48}$$

such that equation (2.46) becomes

$$\begin{aligned} Z(\vartheta) &= \frac{1}{\sqrt{\pi^L}} \int_{\mathbb{R}_+^L} \mathrm{d}^L\mathbf{x} \left\{ \sum_{\mathbf{s}} \prod_{i=1}^{L} \frac{(-1)^{s_i} \exp[-x_i(s_{i-1} + s_i + \vartheta)]}{\sqrt{x_i}} \right\} \\ &= \frac{1}{\sqrt{\pi^L}} \sum_{\mathbf{s}} \prod_{i=1}^{L} \left\{ (-1)^{s_i} \int_0^\infty \mathrm{d}x\, \frac{\exp[-x(s_{i-1} + s_i + \vartheta)]}{\sqrt{x}} \right\} \\ &= \sum_{\mathbf{s}} \prod_{i=1}^{L} \frac{(-1)^{s_i}}{\sqrt{s_{i-1} + s_i + \vartheta}} = \sum_{\mathbf{s}} \prod_{i=1}^{L} T_{s_i s_{i+1}}(\vartheta) \\ &= \mathrm{Tr}\big(T(\vartheta)^L\big) = \lambda_+(\vartheta)^L + \lambda_-(\vartheta)^L \end{aligned} \tag{2.49}$$

with the "transfer matrix"

$$T(\vartheta) = \begin{pmatrix} \frac{1}{\sqrt{\vartheta}} & -\frac{1}{\sqrt{1+\vartheta}} \\ \frac{1}{\sqrt{1+\vartheta}} & -\frac{1}{\sqrt{2+\vartheta}} \end{pmatrix} \tag{2.50}$$

and its eigenvalues

$$\lambda_\pm(\vartheta) = \frac{1}{2\sqrt{\vartheta}} - \frac{1}{2\sqrt{\vartheta+2}} \pm \frac{\sqrt{2\,\vartheta^{3/2}\sqrt{\vartheta+2} - 2\vartheta^2 - 4\vartheta + 2\sqrt{\vartheta\,(\vartheta+2)} + 2}}{2\sqrt{\vartheta\,(\vartheta+1)\,(\vartheta+2)}}\,. \tag{2.51}$$

This and equation (2.47) yields the result for $P_{\max}$ and $\mathbb{E}[h]$. For large $L$, the behavior of $P_{\max}$ is governed by the larger eigenvalue and hence the constant from equation (2.43) is given by

$$\lambda_2 = \lambda_+(1) = \frac{1}{6}\left[3 - \sqrt{3} + \sqrt{6\left(\sqrt{3} - 1\right)}\right] \approx 0.5606\,. \tag{2.52}$$

The expected height reads

$$\mathbb{E}[h] = \frac{L\left[b_- a_-^L + b_+ a_+^L\right]}{12\sqrt{11\sqrt{3}-19}\left[a_-^L + a_+^L\right]}\,, \tag{2.53}$$

where

$$a_\pm = \pm\sqrt{1 \pm 2\sqrt{3\sqrt{3}-5}}$$

and

$$b_\pm = \sqrt{124\sqrt{3} - 209 \pm 2\sqrt{7139\sqrt{3} - 12365}}\,.$$

As expected, the height is asymptotically linear in $L$ with a slope given by

$$\lim_{L\to\infty} \frac{\mathbb{E}[h]}{L} = \frac{b_+}{12\sqrt{11\sqrt{3}-19}} \approx 0.9224\,. \tag{2.54}$$

It should be mentioned for comparison that the fitness of a randomly chosen sequence is given by $\mathbb{E}[h] = L/2$.

# 3. Accessibility Percolation

A path

$$\sigma^1 \to \sigma^2 \to \ldots \to \sigma^n$$

through a fitness landscape is called accessible if the fitness is monotonously increasing along the path [6, 28, 44], i.e., if

$$w\big(\sigma^1\big) < w\big(\sigma^2\big) < \ldots < w(\sigma^n)\,.$$

The importance of these paths is that they can be taken easily by a population since every single step is facilitated by selection. Following a non-accessible path requires the crossing of a fitness valley at some point. This might happen but is usually much more difficult than following an accessible path.

Obviously, the existence of accessibility paths depends on the choice of the start and endpoint. Usually one defines the global maximum $\sigma_{\max}$ as the goal and its antipodal genotype $\overline{\sigma}_{\max}$ as starting point. The quantity of interest is the number $X$ of accessible paths between these genotypes, especially the first moment $\mathbb{E}[X]$ and the probability $\mathbb{P}[X > 0]$ that there is at least one for the limit $L \to \infty$. If needed, additional information (e.g., the dimensionality $L$) will be attached to $X$ as an index (e.g., $X_L$).

Accessibility of a landscape is generally associated with its ruggedness. In a smooth landscape with only one maximum $\sigma_{\max}$ (e. g., the RMF model with $s \to \infty$ or the NK model with $K = 1$) all shortest paths starting from an arbitrary genotype to $\sigma_{\max}$ are accessible. Rugged landscapes, on the other hand, have far less or even no accessible paths. Local maxima, which are another indicator for ruggedness, act as a barrier for accessibility since accessible paths must not contain local maxima by definition (except the final genotype). Therefore, $N_{\max}$ and $X$ are usually negatively correlated.

The study of accessible paths can be interpreted as a certain kind of bond percolation problem (and hence the phrase *accessibility percolation*) on a directed graph. This becomes obvious by replacing all undirected edges of the underlying graph by directed ones pointing towards the genotype with larger fitness. An accessible path in this picture is just a normal path through a directed graph respecting the orientation of edges. The important point is that the orientation is based on the gradient of a globally defined fitness function $w(\sigma)$ rather than being a local property. Therefore, a randomly chosen edge has a probability of $1/2$ to point in either direction, but the orientations of several adjacent edges are correlated.

Moreover, there are also many similarities to *first passage site percolation* [45, 46]. In this type of percolation, a passage time is assigned to each vertex and the question is whether paths between two vertices $\sigma$ and $\tau$ exist such that the total passage time of all vertices along the path (without counting $\sigma$ and $\tau$) is below some threshold. It

can be shown for fitness uniformly distributed on $[0, 1]$ that the probability to find an accessible path from some genotype $\sigma$ with $w(\sigma) = \alpha$ to the global maximum $\sigma_{\max}$ is equal to the probability that there is a path from $\sigma$ to $\sigma_{\max}$ with total passage time smaller than $1 - \alpha$ when the fitness is interpreted as passage time [46]. This result for the percolation probability holds true for arbitrary graphs, but the distribution of the number of paths will differ in general.

It will turn out that the HoC model on the hypercube is in some sense a critical model for accessibility percolation. The probability $\mathbb{P}[X > 0]$ tends to 0 for $L \to \infty$ if only shortest paths are allowed, but slight changes that increase the accessibility will raise the limiting probability to a positive value. This happens, for instance, when also non-shortest paths are allowed or the initial fitness $w(\overline{\sigma}_{\max})$ has a sufficiently small value. In the latter case, $\mathbb{P}[X > 0]$ will even converge to 1. Furthermore, many results suggest that a tree is a good approximation for the hypercube if its parameters are scaled properly. Concerning this scaling, one can observe similar critical behavior: If a regular tree (also known as $n$-ary tree, $n$-tree or Cayley tree) is scaled in a way that the total number of paths is slightly larger than the corresponding number for a hypercube, the percolation probability will tend to 1 while it tends to 0 if the number is slightly smaller.

## 3.1. HoC Model on Trees

Due to their very clear and simple structure, tree graphs are convenient for the study of percolation problems in general and accessibility percolation in particular. Although trees and hypercube graphs might seem to be very different, the local structure of very large hypercubes can be approximated by trees quite well if their shape is chosen properly. In the following, the focus will be on $n$-trees that can be defined recursively as follows (see also figure 3.1(a) for an example): In the first generation $L = 1$, an $n$-tree is a star-like graph consisting of a vertex, called *root* and denoted by $\hat{\mathbf{0}}$ in the following, that is connected to $n$ other vertices which are called *leaves*. Subsequent tree generations $L + 1$ are obtained by connecting each leaf from the $L$-th generation to $n$ new child vertices that become the leaves in this generation. The generation $L$ of the tree will be called *height*, the number $n$ of children per vertex will be called *branching number*. There are $n^L$ leaves in the tree, each of them corresponds uniquely to a path consisting of $L+1$ vertices from the root to the leaf. This means that such a path has the same length as a path in the $L$-dimensional hypercube between antipodal vertices. $X_L$ will denote the number of accessible paths from the root to the leaf in the context of trees. The branching number $n$ will be scaled later to make the tree mimic a hypercube structure.

Fitness landscape models of HoC (and RMF) type are straightforward to generalize for the usage on a tree: The analogue of the HoC model is to assign independent random numbers $w(\sigma)$ to each vertex $\sigma$ drawn from an arbitrary continuous distribution. Since the destination vertex is usually assumed to be the global optimum on the hypercube, it will be assumed analogously that the root is the global minimum of the tree. Note that this assumption is made to emphasize the similarity between trees and hypercubes and without loss of generality due to the recursive structure of trees (see next section).

### 3.1.1. Recurrence Relation for the Percolation Probability

It should be obvious from the definition of the tree that it has a recursive structure, i.e., each vertex $\sigma$ adjacent to the root $\hat{\mathbf{0}}$ of a tree $T(\hat{\mathbf{0}})$ with height $L$ can be interpreted as the root of a sub-tree $T(\sigma)$ of height $L-1$. The original tree $T(\hat{\mathbf{0}})$ is accessible if and only if there is at least one vertex $\sigma$ adjacent to $\hat{\mathbf{0}}$ such that $w(\hat{\mathbf{0}}) < w(\sigma)$ and $T(\sigma)$ is an accessible sub-tree. Let

$$Q_L(x) = \mathbb{P}\big[X_L = 0 \mid w(\hat{\mathbf{0}}) = x\big]$$

be the probability that there are no accessible paths in a tree, given that the root has fitness $x$. Then $Q_L(x)$ obeys the recurrence relation

$$Q_{L+1}(x) = \left[F_L(x) + \int_x^\infty f_L(y) Q_L(y)\,\mathrm{d}y\right]^n \tag{3.1}$$

with $Q_1(x) = F_0(x)^n$. $F_d(x)$ is the cumulative distribution function of $w(\sigma)$ for vertices $\sigma$ that are at distance $d$ to the closest leaf and $f_d(x)$ is the corresponding probability distribution. Note that the explicit $d$-dependence can be omitted in the HoC case, but is needed for the RMF model that will be studied later. The probability of having at least one accessible path in a tree of height $L$ is then given by

$$\mathbb{P}[X_L > 0] = 1 - \int_{-\infty}^\infty f_L(x) Q_L(x)\,\mathrm{d}x = 1 - \left[\lim_{x\to-\infty} Q_{L+1}(x)\right]^{\frac{1}{n}} \tag{3.2}$$

if the fitness $w(\hat{\mathbf{0}})$ of the root is drawn randomly and by

$$\mathbb{P}[X_L > 0] = 1 - \lim_{x\to-\infty} Q_L(x) \tag{3.3}$$

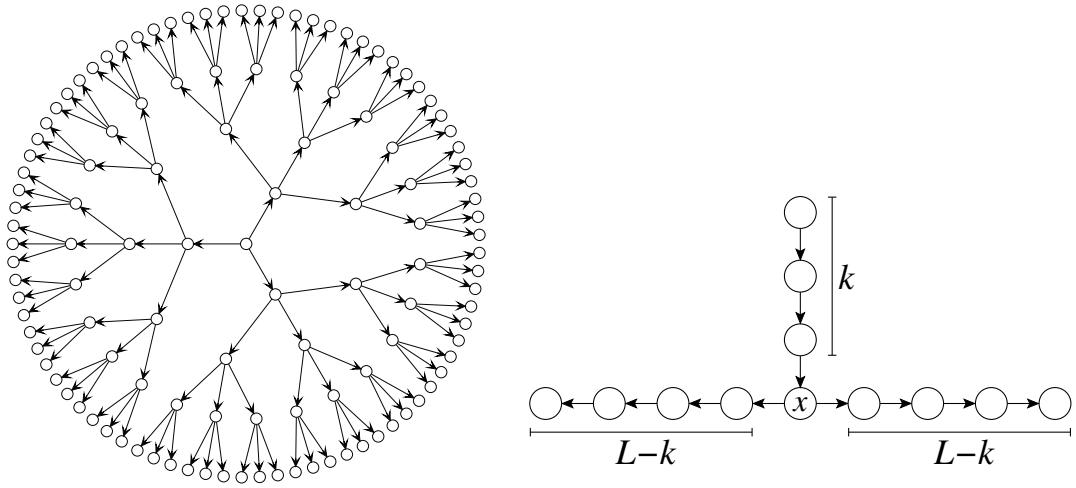
in case the root's fitness is the global minimum, which is assumed for the HoC model. Unfortunately, there is no expression for $Q_L(x)$ in simple closed form, but the integral in equation (3.1) can be approximated numerically in order to gain supplementary information to the analytical study.

### 3.1.2. First and Second Moment of $X_L$ for the HoC Model

It will be useful in the following to define indicator variables for each path by

$$\theta_i = \begin{cases} 1 & \text{if the } i\text{-th path is accessible,} \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

In the HoC model, the fitness distribution does not have an influence on the accessibility. Therefore, a uniform distribution on $[0,1]$ will be assumed in the following for simplicity. A path from the root to a leaf consists of $L+1$ vertices. Taking into account that the root is assumed to be the global minimum, there are $L!$ possible permutations of the remaining fitness values. All of them have equal probability to occur but only one




**(a)** Example tree **(b)** Two paths with T-junction

**Figure 3.1.** (a) Example of a regular tree with height $L = 4$ and branching number $n = 3$. The root is the vertex in the center. (b) Two paths through a tree that share $k + 1$ common vertices.

corresponds to an accessible path. Therefore, the probability that $\theta_i = 1$ is given by $1/L!$ and hence $\mathbb{E}[\theta_i^m] = 1/L!$ for any $m > 0$. For the first two moment of $X_L$, this yields

$$\mathbb{E}[X_L] = \mathbb{E}\left[\sum_{i=1}^{n^L} \theta_i\right] = \sum_{i=1}^{n^L} \mathbb{E}[\theta_i] = \frac{n^L}{L!} \tag{3.5}$$

and

$$\mathbb{E}\big[X_L^2\big] = \sum_{i=1}^{n^L}\sum_{j=1}^{n^L} \mathbb{E}[\theta_i\theta_j] = \mathbb{E}[X_L] + 2\sum_{i=1}^{n^L}\sum_{j=i+1}^{n^L} \mathbb{E}[\theta_i\theta_j]\,. \tag{3.6}$$

The correlator $\mathbb{E}[\theta_i\theta_j]$ is nothing but the probability that both paths $i$ and $j$ are accessible which, in turn, depends only on the number of vertices both paths share. Suppose a pair of paths as shown in figure 3.1(b), i.e., the vertex $\sigma$ where both paths separate has fitness $w(\sigma) = x$ and is the $(k+1)$-th vertex in the path. Then both paths are accessible if and only if all vertices closer to the root than $\sigma$ have a fitness smaller than $x$, all vertices that are further away have a fitness larger than $x$ and the fitness along each of the three sub-paths separated by $\sigma$ has to be in ascending order. Combining these conditions and integrating over $x$, the probability $p_k$ that both paths are accessible is given by

$$p_k = \int_0^1 \frac{x^{k-1}}{(k-1)!}\left(\frac{(1-x)^{L-k}}{(L-k)!}\right)^2 = \binom{2L-2k}{L-k}\frac{1}{(2L-k)!}\,. \tag{3.7}$$

In order to evaluate the sum in equation (3.6) one needs to know the number $m_k$ of such pairs of paths, too. The number of (ordered) pairs with $k+1$ common vertices can be counted as follows: Two paths in a tree correspond uniquely to two leaves. The first leaf can be chosen arbitrarily among all $n^L$ leaves. To reach the second leaf, one follows the first path for $k+1$ vertices, then there are $n-1$ possible new vertices and finally there are $L-k-1$ additional branching points with $n$ possible vertices each. In total, there are

$$m_k = (n-1)\, n^{2L-k-1}$$

pairs of paths with $k+1$ common vertices. With this, equation (3.6) can finally be written as

$$\begin{aligned}\mathbb{E}\big[X_L^2\big] &= \mathbb{E}[X_L] + \sum_{k=0}^{L-1} p_k\, m_k = \mathbb{E}[X_L] + n^{2L}\,\frac{(n-1)}{n}\sum_{k=0}^{L-1}\binom{2L-2k}{L-k}\,\frac{n^{-k}}{(2L-k)!} \\ &= \mathbb{E}[X_L] + n^L\,\frac{(n-1)}{n}\sum_{k=1}^{L}\binom{2k}{k}\,\frac{n^k}{(L+k)!}\,. \end{aligned} \tag{3.8}$$

### 3.1.3. Probability of having Accessible Paths

Here and in following sections, the probability $\mathbb{P}[X_L]$ will be estimated by inequalities based on the first and second moment. Let $Z$ be a non-negative random variable with finite second moment. According to Markov's inequality, the first moment $\mathbb{E}[Z]$ is an upper limit for the probability $\mathbb{P}[Z>0]$. Since $\mathbb{E}[X_L] = n^L/L! \to 0$ for $L\to\infty$ and any fixed $n\in\mathbb{N}$, the probability for having at least one accessible path vanishes in the limit of an infinite tree. A lower limit of $\mathbb{P}[Z>0]$ follows from the Cauchy-Schwarz inequality since

$$\mathbb{E}[Z]^2 = \mathbb{E}\big[Z\, I_{[Z>0]}\big]^2 \le \mathbb{E}\big[Z^2\big]\,\mathbb{E}\Big[I_{[Z>0]}^2\Big] = \mathbb{E}\big[Z^2\big]\,\mathbb{P}[Z>0] \tag{3.9}$$

and hence

$$\mathbb{P}[Z>0] \ge \frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}\,, \tag{3.10}$$

where $I_{[A]}$ is the indicator variable of the event $A$. Using this inequality, it will be shown in the following that $\mathbb{P}[X_L>0]$ is asymptotically equivalent to the mean value $\mathbb{E}[X_L]$ if $n$ is kept constant (see also figure 3.2). The following bounds on Stirling's formula

$$\sqrt{2\pi}\; N^{N+1/2}e^{-N} \le N! \le e\; N^{N+1/2}e^{-N} \tag{3.11}$$

applied to equation (3.8) yield

$$\begin{aligned}
\mathbb{E}\big[X_L^2\big] &\leq \mathbb{E}[X_L] + \mathbb{E}[X_L]^2 + n^L \sum_{k=1}^{L-1} \binom{2k}{k} \frac{n^k}{(L+k)!} \\
&\leq \mathbb{E}[X_L] + \mathbb{E}[X_L]^2 + n^L \sum_{k=1}^{L-1} \frac{2^{2k-1} e^{k+L+1} n^k}{\pi^{3/2}\sqrt{k}(k+L)^{k+L+\frac{1}{2}}} \\
&\leq \mathbb{E}[X_L] + \mathbb{E}[X_L]^2 + \frac{n^L}{\sqrt{2\pi}\ L^{L+1/2} e^{-L}} \sum_{k=1}^{L-1} \left(\frac{4en}{L}\right)^k \\
&\leq \mathbb{E}[X_L]^2 + \mathbb{E}[X_L] \left[1 + \left(\frac{4en}{L}\right) \frac{1 - \left(\frac{4en}{L}\right)^{L-1}}{1 - \left(\frac{4en}{L}\right)}\right] \\
&\leq \mathbb{E}[X_L]^2 + \mathbb{E}[X_L] \Big(1 + \frac{c\,n}{L}\Big) \qquad (3.12)
\end{aligned}$$

for some positive constant $c$ and sufficiently large $L$. Using the first moment and equation (3.10) as upper and lower bound, respectively, results in

$$\mathbb{E}[X_L] \geq \mathbb{P}[X_L > 0] \geq \frac{\mathbb{E}[X_L]}{1 + \mathbb{E}[X_L] + \frac{c\,n}{L}} \qquad (3.13)$$

and hence

$$\mathbb{P}[X_L > 0] \sim \mathbb{E}[X_L]\,, \qquad (3.14)$$

where the tilde means that both quantities are asymptotically equivalent, i.e., $\lim_{L\to\infty} \mathbb{P}[X_L > 0]/\mathbb{E}[X_L] = 1$.

### 3.1.4. Scaling of the Branching Number

Treating the branching number as a function of the height, i.e., $n = n(L)$, leads to interesting consequences. To begin with, Stirling's approximation applied to the mean value of $X_L$ yields

$$\mathbb{E}[X_L] = \frac{n(L)^L}{L!} \sim \frac{(e\,n(L)/L)^L}{\sqrt{2\pi L}}\,. \qquad (3.15)$$

This means that for $L \to \infty$ and $n/L \to \alpha$, there are on average infinitely many paths if $\alpha > e^{-1}$ and no paths if $\alpha < e^{-1}$. In the latter case, also the probability $\mathbb{P}[X_L > 0]$ goes to zero since $\mathbb{E}[X_L]$ serves as an upper limit. In case $n/L \to 0$, also the previously obtained equation (3.12) and in turn equations (3.13) and (3.14) hold true, i.e., the probability to have accessible path is still asymptotically equivalent to the mean of $X_L$ that goes to zero. As mentioned before, the scaling is intended to mimic the hypercube. Demanding that the total number of paths through tree and hypercube are equal yields

$$n^L \stackrel{!}{=} L! \quad \Rightarrow \quad n = \sqrt[L]{L!} \sim \frac{L}{e} \cdot (2\pi L)^{1/2L} \sim \frac{L}{e}\,, \qquad (3.16)$$

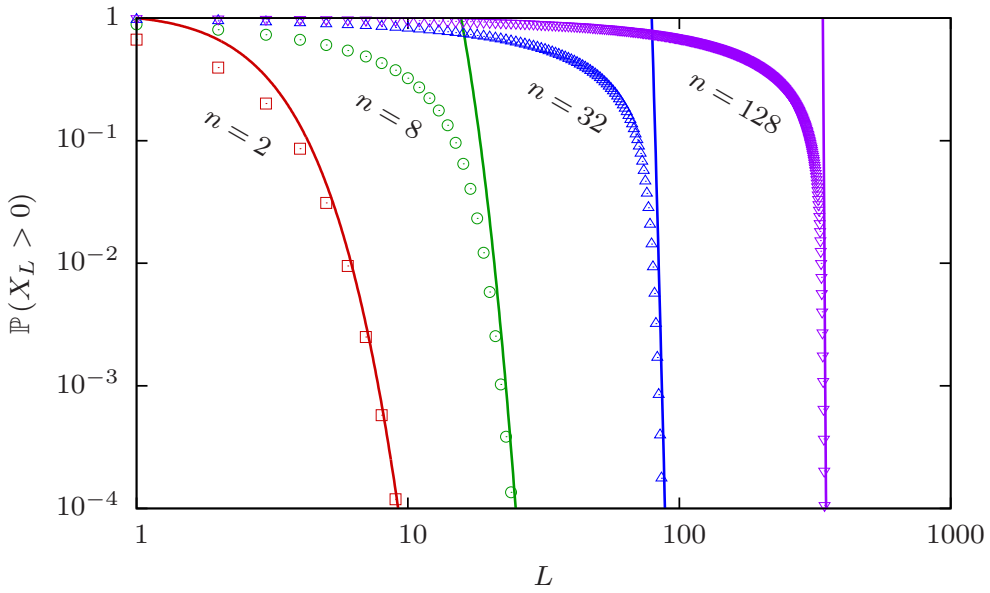



**Figure 3.2.** Probability $\mathbb{P}[X_L > 0]$ of having at least one accessible path in a regular tree. Symbols correspond to the numerical solution of equation (3.3), lines correspond to the asymptotic value $\mathbb{E}[X_L] = n^L/L!$.

which corresponds asymptotically to the threshold where $\mathbb{E}[X_L]$ jumps from zero to infinity.

The case where $n = \alpha L$ and $\alpha > e^{-1}$ will be considered in the following. Combining equations (3.8) and (3.10) yields

$$\mathbb{P}[X_L > 0] \geq \frac{\mathbb{E}[X_L]^2}{\mathbb{E}\left[X_L^2\right]} \geq \frac{1}{1/\mathbb{E}[X_L] + 1 + S(L)}\,, \tag{3.17}$$

where

$$S(L) = \frac{L!^2}{n^L} \sum_{k=1}^{L-1} \binom{2k}{k} \frac{n^k}{(L+k)!}\,. \tag{3.18}$$

Following [47], the function $S(L)$ can be estimated recursively which reads

$$\begin{aligned} S(L+1) &= \frac{(L+1)!^2}{n^{L+1}} \sum_{k=1}^{L} \binom{2k}{k} \frac{n^k}{(L+k+1)!} \\ &\leq \frac{(L+1)^2}{n\,(2L+1)} + \frac{(L+1)^2}{n\,(L+2)}\, S(L) \leq \frac{1+S(L)}{\alpha}\,. \end{aligned} \tag{3.19}$$

With this, one can show by induction that $S(L) < 1/(\alpha - 1)$ for $\alpha > 1$ and hence $\lim_{L\to\infty} \mathbb{P}[X_L > 0] > 0$ according to (3.17). Unfortunately, this methods fails for $\alpha \in (e^{-1}, 1]$.

An improvement of the lower bound was accomplished in [48]. The strategy is to introduce paths that are accessible and fulfill additionally the condition that the $i$-th vertex $\sigma_i$ after the root has fitness of at least $w(\sigma_i) > \epsilon + (1-\epsilon)(i-1)/L$. If the number of these paths is denoted by $X_{L,\epsilon}$ and $\alpha > e^{-1}$, it can be shown via inequality (3.10) that $\mathbb{P}[X_{L,\epsilon}] > c/L^3$, where $c$ is some positive constant. Then one can argue that the number of paths contributing to $X_{L,\epsilon}$ up to the fourth level is at least of order $L^4$ which compensates for the $1/L^3$ decay obtained from the moment analysis such that $\mathbb{P}[X_{L,\epsilon} > 0] \to 1$ as $L \to \infty$. Since $X_{L,\epsilon} < X_L$, this also proves the phase transition

$$\lim_{L\to\infty} \mathbb{P}[X_L > 0] = \begin{cases} 0 & \text{if} \quad \alpha < e^{-1}\,, \\ 1 & \text{if} \quad \alpha > e^{-1}\,. \end{cases} \tag{3.20}$$

The complete proof is presented in appendix A.3.

### 3.1.5. Further Results

Concerning the distribution of $X_L$, it was shown that for $n = L$ and the root having fitness $w(\hat{\mathbf{0}}) = \alpha/L$, the rescaled variable $L!\, X_L/L^L$ converges in law to an exponential distribution with mean value $\exp(-\alpha)$ [49]. Higher orders of the critical scaling of the branching number $n$ were obtained in [50]. For $n = e\,L - \beta\,\log(L)$, the threshold where $\lim_{L\to\infty} \mathbb{P}[X_L > 0]$ jumps from 0 to 1 is at $\beta = 3/2$.

Also other types of spherically symmetric trees were studied, i.e., trees with a branching number $n = n(d)$ depending on the distance $d$ to the root. The case $n(d) = \lceil \alpha\,(d+1) \rceil$, i.e., a tree whose branching number increases with distance to the root was studied in [51]. Results are that the percolation probability $\mathbb{P}[X_L > 0]$ converges to zero for $\alpha \leq 1$ and $\lim_{L\to\infty} \mathbb{P}[X_L > 0] > 0$ for $\alpha > 1$. The case $n(d) = L - d$, i.e., a diminishing tree, where each leaf $\sigma$ has fitness $w(\sigma) = 1$, was studied in [52]. It was found that for a root fitness of $w(\hat{\mathbf{0}}) = \alpha/L$, the variable $X_L/L$ converges in law to an exponential distribution with mean value $\exp(-\alpha)$, like in the case described above. Moreover, if $w(\hat{\mathbf{0}})$ is chosen at random, one has $\mathbb{P}[X_L > 0] \sim \log(L)/L$, which reflects the behavior on the hypercube.

## 3.2. HoC Model on the Directed Hypercubes

In this section, the objects of interest are accessible paths to the global maximum $\sigma_{\max}$ of a directed hypercube that start from the antipodal sequence $\overline{\sigma}_{\max}$. The adjective "directed" refers to the fact that only shortest paths are considered, i.e., all edges are directed towards $\sigma_{\max}$. Without loss of generality, the sequence $\sigma_{\max} = \hat{\mathbf{1}} = (1, \ldots, 1)$ denotes the global fitness maximum and hence $\hat{\mathbf{0}} = (0, \ldots, 0)$ is the starting sequence of all paths under consideration. Like before, the fitness distribution does not influence the results on the HoC landscape and hence it is assumed that all fitness values are uniformly distributed on $[0, 1]$, except of $w(\hat{\mathbf{1}}) = 1$. Later on, it will be useful to constrain the starting vertex to have a specific fitness value $w(\hat{\mathbf{0}}) = \alpha$. This model is known as the $\alpha$-constrained HoC model [18] (or short: $\alpha$-CHoC). A version where $\alpha = 0$ was

previously studied in [28, 53]. Analogously to the notation for the tree, let $X_L$ and $X_{L,\alpha}$ be the number of accessible paths in the standard HoC and $\alpha$-CHoC model, respectively. The main result, which was obtained by Hegarty and Martinsson [18], is

$$\lim_{L\to\infty} \mathbb{P}[X_{L,\alpha} > 0] = \begin{cases} 0 & \text{if } \alpha = (\log L + \epsilon_L)/L\,, \\ 1 & \text{if } \alpha = (\log L - \epsilon_L)/L\,, \end{cases} \tag{3.21}$$

for any sequence $\epsilon_L \to \infty$ with $0 < \epsilon_L < \log L$. As a corollary it follows that

$$\mathbb{P}[X_L > 0] \sim \frac{\log L}{L}\,. \tag{3.22}$$

The proof is again based on the moments of $X_L$ and will be somewhat similar to the computation of the corresponding quantities for the tree.

### 3.2.1. First Moment and Upper Limits for Accessible Paths

Consider the standard HoC model. There are $L!$ shortest path from $\hat{\mathbf{0}}$ to $\hat{\mathbf{1}}$. Each of them has a probability of $1/L!$ to be accessible and hence $\mathbb{E}[X_L] = 1$. Unlike for the tree, the mean is obviously not a useful upper bound for the probability $\mathbb{P}[X_L > 0]$ to have at least one accessible path. However, by conditioning the fitness of the starting sequence, the bound can be improved. If $w(\hat{\mathbf{0}}) > \alpha$, all subsequent fitness values have to be larger than $\alpha$ as well. Therefore, the probability that a randomly chosen path is accessible, conditioned on $w(\hat{\mathbf{0}}) > \alpha$, is given by $(1-\alpha)^{L-1}/L!$ and hence

$$\begin{aligned} \mathbb{P}[X_L > 0] &= \mathbb{P}\big[X_L > 0 \wedge w(\hat{\mathbf{0}}) < \alpha\big] + \mathbb{P}\big[X_L > 0 \wedge w(\hat{\mathbf{0}}) > \alpha\big] \\ &\le \mathbb{P}\big[w(\hat{\mathbf{0}}) < \alpha\big] + \mathbb{E}\big[X_L \mid w(\hat{\mathbf{0}}) > \alpha\big]\, \mathbb{P}\big[w(\hat{\mathbf{0}}) > \alpha\big] = \alpha + (1-\alpha)^L\,. \end{aligned} \tag{3.23}$$

Since $\alpha$ is arbitrary, one may chose the value

$$\alpha = 1 - \left(\frac{1}{L}\right)^{\frac{1}{L-1}} = 1 - \exp\left(-\frac{\log L}{L-1}\right) = \frac{\log L}{L} + \mathcal{O}\left(\left(\frac{\log L}{L}\right)^2\right) \tag{3.24}$$

which minimizes the right-hand side of equation (3.23). Then the upper limit becomes

$$\mathbb{P}[X_L > 0] \le 1 - \left(\frac{1}{L}\right)^{\frac{1}{L-1}} + \left(\frac{1}{L}\right)^{\frac{L}{L-1}} = \frac{\log L}{L} + \mathcal{O}\left(\frac{1}{L}\right). \tag{3.25}$$

The fact that this is a sharp bound and hence equation (3.22) holds true will follow as a corollary from the study of the $\alpha$-CHoC model which, in turn, is based on the study of the moments of $X_{L,\alpha}$.

Consider the $\alpha$-CHoC case in the following, i.e., it is assumed that $w(\hat{\mathbf{0}}) = \alpha$. The probability that a specific path is accessible is given by $(1-\alpha)^{L-1}/(L-1)!$ and hence the mean value of $X_{L,\alpha}$ is given by

$$\mathbb{E}[X_{L,\alpha}] = L\,(1-\alpha)^{L-1}\,. \tag{3.26}$$

For $\alpha = (\log L + \epsilon_L)/L$ with $0 < \epsilon_L < \log L$ and $\epsilon_L \to \infty$ one has

$$\mathbb{E}[X_{L,\alpha}] = L\left(1 - \frac{\log L + \epsilon_L}{L}\right)^{L-1} \leq Le^{-(\log L + \epsilon L)} = e^{-\epsilon_L} \xrightarrow{L\to\infty} 0\,. \tag{3.27}$$

This already proves the first half of equation (3.21). Conversely, if $\alpha = (\log L - \epsilon_L)/L$ one has

$$\begin{aligned}\mathbb{E}[X_{L,\alpha}] &= L\left(1 - \frac{\log L - \epsilon_L}{L}\right)^{L-1} \geq L\left[\left(1 - \frac{\log L}{L}\right)\cdot\left(1 + \frac{\epsilon_L}{L}\right)\right]^{L-1} \\ &\geq L\left[\left(1 - \frac{\epsilon_L^2}{L^2}\right)\right]^{L-1} \geq L\left[\left(1 - \frac{1}{L}\right)\right]^{L-1} \xrightarrow{L\to\infty} \infty\,.\end{aligned} \tag{3.28}$$

For the proof that also $\mathbb{P}[X_{L,\alpha} > 0] \to 1$ in this case, the second moment will be needed as well.

### 3.2.2. Second Moment and Lower Limits for Accessible Paths

In order to study the lower limit for $\mathbb{P}[X_{L,\alpha} > 0]$, let $\alpha = (\log L - \epsilon_L)/L$ in the following such that $\mathbb{E}[X_{L,\alpha}] \to \infty$. The calculation of the second moment follows the same procedure as before: Let $\theta_i$ be the indicator variable that the $i$-th path is accessible, then

$$\mathbb{E}\big[X_{L,\alpha}^2\big] = \sum_{i,j} \mathbb{E}[\theta_i\theta_j]\,. \tag{3.29}$$

Like on the tree, $\mathbb{E}[\theta_i\theta_j]$ is nothing but the probability that both paths $i$ and $j$ are accessible. But unlike on the tree, the probability does not only depend on the number of common vertices, since the paths might diverge and converge several times until the final vertex $\hat{\mathbf{1}}$ is reached (see figure 3.3 for an example). Suppose path $i$ and $j$ have $k-1$ common interior vertices, i.e., $2L-k-1$ vertices in total (aside from $\hat{\mathbf{0}}$ and $\hat{\mathbf{1}}$). The fitness of each vertices is a random number and hence there are $(2L-k-1)!$ different possibilities to order the fitness ranks among these vertices. How many possibilities lead to the accessibility of both paths? The smallest numbers have to be put in the beginning before $i$ and $j$ diverge for the first time, while the largest numbers have to be put after the last time both paths merge. Only in between is some freedom to distribute the fitness values. Suppose all $2L-2k$ separated vertices are consecutive along the paths, i.e., the paths diverge and merge exactly once. Then one can put any subset of size $L-k$ of random numbers to path $i$ and the remaining ones on path $j$ which gives $\binom{2L-2k}{L-k}$ possibilities. For paths that diverge and merge several times, there are less possibilities. Taking into account that all fitness values must additionally be larger than $\alpha$, an upper limit for $\mathbb{E}[\theta_i\theta_j]$ is given by

$$\mathbb{E}[\theta_i\theta_j] \leq \binom{2L-2k}{L-k}\frac{(1-\alpha)^{2L-k-1}}{(2L-k-1)!} \tag{3.30}$$

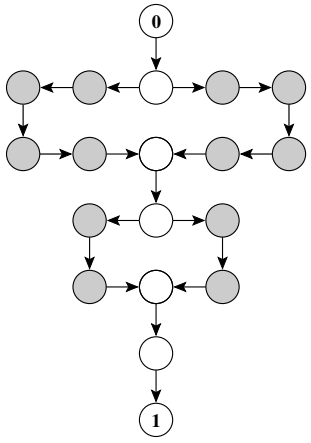



**Figure 3.3.** Example of a pair of paths through an $L = 12$ hypercube with 5 common interior vertices (white circles) and 6 separated vertices (grey circles) for each path.

with equality if path $i$ and $j$ diverge and merge exactly once. Not surprisingly, this is the same as equation (3.7) apart from the constraint $w(\hat{\mathbf{0}}) = \alpha$.

In order to evaluate the sum (3.29), one also needs to know the number of pairs of paths that share $k-1$ interior vertices. As mentioned in section 2.1, this number is given by $L! \cdot T(L,k)$ where $T(L,k)$ is the number of permutations with $k$ *components* of integer numbers from 1 to $L$. This yields

$$\mathbb{E}\big[X_{L,\alpha}^2\big] \leq \sum_{k=1}^{L} L! \cdot T(L,k) \cdot \binom{2L-2k}{L-k} \frac{(1-\alpha)^{2L-k-1}}{(2L-k-1)!}\,. \tag{3.31}$$

Hegarty and Martinsson provided several useful formulas for $T(L,k)$ [18]. In the following, one will need the following two inequalities: For $0 < \delta < 1$ there is a constant $C_\delta$ such that

$$T(L,k) \leq C_\delta\, k\, (L-k+1)! \tag{3.32}$$

for $0 < k \leq (1-\delta)L$ and there is a constant c such that

$$T(L,k) \leq c\,(L-k+1) \cdot \left(\frac{3L-2k}{5}\right)^k \tag{3.33}$$

for all $0 < k \leq L$. Note that the inequalities (3.32) and (3.33) are more useful for small and large $k$, respectively. This is why the sum (3.29) will be split into two parts which

reads

$$\mathbb{E}\big[X_{L,\alpha}^2\big] \leq \underbrace{\sum_{k=1}^{(1-\delta)L} L! \cdot T(L,k) \cdot \binom{2L-2k}{L-k} \frac{(1-\alpha)^{2L-k-1}}{(2L-k-1)!}}_{=S_1} + \underbrace{\sum_{k=0}^{\delta L} L! \cdot T(L,L-k) \cdot \binom{2k}{k} \frac{(1-\alpha)^{L+k-1}}{(L+k-1)!}}_{=S_2} . \tag{3.34}$$

Both parts $S_1$ and $S_2$ will be bounded from above in the following. Applying inequality (3.32), Stirling's formula and the fact that $k < (1-\delta)L$ to $S_1$ yields

$$\begin{aligned}
&\binom{2L-2k}{L-k} \frac{L! \cdot T(L,k)}{(2L-k-1)!} \leq \binom{2L-2k}{L-k} \frac{C_\delta\, L!\, k\, (L-k+1)!}{(2L-k-1)!} \\
&\leq \frac{e^3\, C_\delta}{2\pi^{3/2}} \sqrt{L\,(2L-k)}\, k\, (L-k+1)\, \frac{L^L\, 4^{L-k}\, (L-k)^{L-k}}{(2L-k)^{2L-k}} \\
&\leq \frac{e^3\, C_\delta}{\sqrt{2\pi^3}}\, L^2\, k\, \frac{L^L\, 4^{L-k}\, (L-k)^{L-k}}{(2L-k)^{2L-k}} = C'_\delta\, L^2\, k\, 2^{-k} \left(\frac{L-k}{L}\right)^{L-k} \left(\frac{2L-k}{2L}\right)^{k-2L} \\
&\leq C'_\delta\, L^2\, k\, 2^{-k} \left(\frac{2L-k}{2L}\right)^{2L-2k} \left(\frac{2L-k}{2L}\right)^{k-2L} = C'_\delta\, L^2\, k\, 2^{-k} \left(\frac{2L}{2L-k}\right)^{k} \\
&\leq C'_\delta\, L^2\, k\, 2^{-k} \left(\frac{2L}{2L-(1-\delta)L}\right)^{k} = C'_\delta\, \frac{L^2\, k}{(1+\delta)^k} ,
\end{aligned}$$

where $C'_\delta = e^3\, C_\delta / \sqrt{2\pi^3}$. This leads to

$$S_1 \leq C'_\delta L^2\, (1-\alpha)^{2L-1} \sum_{k=1}^{(1-\delta)L} \frac{k}{[(1+\delta)\,(1-\alpha)]^k} . \tag{3.35}$$

Now keep in mind that $\alpha = \alpha_L$ is a sequence converging to zero such that $(1+\delta)\,(1-\alpha) > (1+\delta/2)$ for sufficiently large $L$. Then

$$\begin{aligned}
S_1 &\leq C'_\delta L^2\, (1-\alpha)^{2L-1} \sum_{k=1}^{\infty} \frac{k}{\left(1+\frac{\delta}{2}\right)^k} \\
&= C'_\delta L^2\, (1-\alpha)^{2L-1}\, \frac{4+2\delta}{\delta^2} = \tilde{C}_\delta\, L^2\, (1-\alpha)^{2L-1} \\
&= \tilde{C}_\delta\, (1-\alpha)\, \mathbb{E}[X_{L,\alpha}]^2 ,
\end{aligned} \tag{3.36}$$

where the constants involving $\delta$ are absorbed into $\tilde{C}_\delta$ in the last steps.

Consider the second part of (3.34) in the following. Using inequality (3.33) leads to

$$\begin{aligned}
S_2 &= \sum_{k=0}^{\delta L} L! \cdot T(L, L-k) \cdot \binom{2k}{k} \frac{(1-\alpha)^{L+k-1}}{(L+k-1)!} \\
&\leq c \sum_{k=0}^{\delta L} (k+1) \left(\frac{L+2k}{5}\right)^k \binom{2k}{k} \frac{L!}{(L+k-1)!} (1-\alpha)^{L+k-1} \\
&\leq c(1-\alpha)^{L-1} \sum_{k=0}^{\delta L} \frac{k+1}{L^{k-1}} \left(\frac{L+2k}{5}\right)^k 4^k \\
&= cL(1-\alpha)^{L-1} \sum_{k=0}^{\delta L} (k+1) \left(\frac{4L+8k}{5\,L}\right)^k \\
&\leq cL(1-\alpha)^{L-1} \sum_{k=0}^{\delta L} (k+1) \left(\frac{4+8\delta}{5}\right)^k .
\end{aligned}$$

Assuming $\delta < 1/8$, the sum converges for $L \to \infty$ to yet another constant $c_\delta$ and hence

$$S_2 \leq c_\delta L(1-\alpha)^{L-1} = c_\delta \, \mathbb{E}[X_{L,\alpha}] \, . \tag{3.37}$$

Now it is time to combine the estimates for $S_1$ and $S_2$ to get a first bound on the probability to have accessible paths via inequality (3.10):

$$\mathbb{P}[X_{L,\alpha} > 0] \geq \frac{\mathbb{E}[X_{L,\alpha}]^2}{\mathbb{E}\left[X_{L,\alpha}^2\right]} \geq \frac{\mathbb{E}[X_{L,\alpha}]^2}{\tilde{C}_\delta \, (1-\alpha) \, \mathbb{E}[X_{L,\alpha}]^2 + c_\delta \, \mathbb{E}[X_{L,\alpha}]} \, . \tag{3.38}$$

As shown before, $\mathbb{E}[X_{L,\alpha}] \to \infty$ for the choice of $\alpha$ used here which yields

$$\lim_{L\to\infty} \mathbb{P}[X_{L,\alpha} > 0] \geq \frac{1}{\tilde{C}_\delta} \, , \tag{3.39}$$

i.e., the probability to have accessible paths converges to a positive constant. This estimate could in principle be improved to a constant that is at least $1/4$ [18], but in the next section it will be shown that any positive constant is sufficient to raise the limit to 1.

### 3.2.3. Improving the Lower Bound on $\mathbb{P}[X_{L,\alpha} > 0]$

Let $\hat{\mathbf{0}}_i$ and $\hat{\mathbf{1}}_i$ denote the sequence consisting only of zeroes and ones, respectively, except for the $i$-th locus, i.e., $d(\hat{\mathbf{0}}_i, \hat{\mathbf{0}}) = d(\hat{\mathbf{1}}_i, \hat{\mathbf{1}}) = 1$. The first goal is to find four distinct integers $i_1, i_2, j_1$ and $j_2$ such that

$$\alpha < w(\hat{\mathbf{0}}_{i_1}), w(\hat{\mathbf{0}}_{i_2}) \leq \alpha + \frac{\epsilon_L}{3L} \qquad \text{and} \qquad w(\hat{\mathbf{1}}_{j_1}), w(\hat{\mathbf{1}}_{j_2}) \geq 1 - \frac{\epsilon_L}{3L} \, . \tag{3.40}$$

Consider the probability $q_L$ to find these numbers. The number of vertices with fitness in $[\alpha, \alpha + \epsilon_L/3L]$ can be interpreted as the sum of $L$ Bernoulli distributed random

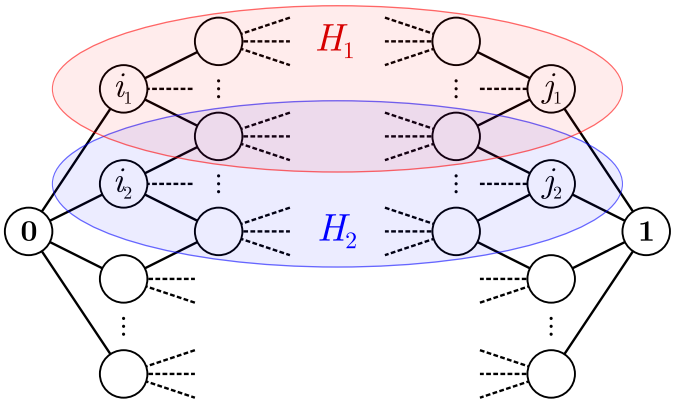



**Figure 3.4.** Construction to improve the lower bound of $\mathbb{P}[X_{L,\alpha} > 0]$. See the text for more details.

variables that take the value 1 with probability $\epsilon_L/3L$. Given that $i_1$ and $i_2$ are already determined, the number of possible vertices with fitness larger than $1 - \epsilon_L/3L$ is the sum of $L - 2$ Bernoulli random variables with the same probability. A version of the Chernoff bound [54] states that

$$\mathbb{P}\left[Z < \frac{\mathbb{E}[Z]}{2}\right] \leq \exp\left(-\frac{\mathbb{E}[Z]}{8}\right), \tag{3.41}$$

where $Z$ is the sum of Bernoulli distributed random variables. This inequality applied to $q_L$ yields

$$q_L > 1 - \exp\left(-\frac{\epsilon_L}{24} - \frac{(L-2)\cdot\epsilon_L}{24L}\right). \tag{3.42}$$

Once those vertices that fulfill (3.40) are found, one can estimate the probabilities to find accessible paths from $\hat{\mathbf{0}}_{i_1}$ to $\hat{\mathbf{1}}_{j_1}$ or from $\hat{\mathbf{0}}_{i_2}$ to $\hat{\mathbf{1}}_{j_2}$. Note that by construction any of these accessible paths can be connected to $\hat{\mathbf{0}}$ and $\hat{\mathbf{1}}$ as well. In order to find a path from $\hat{\mathbf{0}}_{i_k}$ to $\hat{\mathbf{1}}_{j_k}$, all loci but $i_k$ and $j_k$ are relevant, i.e., the underlying structure is a hypercube $H_k$ of dimension $L - 2$ (see figure 3.4) and hence

$$\mathbb{P}\big[\text{Accessible path from } \hat{\mathbf{0}}_{i_1} \text{ to } \hat{\mathbf{1}}_{j_1}\big] = \mathbb{P}[X_{L-2,\tilde{\alpha}} > 0], \tag{3.43}$$

where $\tilde{\alpha} = \alpha + 2\epsilon_L/3L = (\log L - \epsilon_L/3)/L$. The same applies for $i_2$ and $j_2$, but since the hypercube $H_1$ overlaps with $H_2$, the probabilities are not independent. However, an accessible path in $H_2$ is not contained in $H_1$ if the $j_1$-th locus is flipped before the $i_1$-th locus, that is with probability 1/2, and hence

$$\mathbb{P}\big[\text{Accessible path not contained in } H_1 \text{ from } \hat{\mathbf{0}}_{i_2} \text{ to } \hat{\mathbf{1}}_{j_2}\big] \geq \frac{\mathbb{P}[X_{L-2,\tilde{\alpha}} > 0]}{2}. \tag{3.44}$$

Since an accessible path in either $H_1$ or $H_2 \setminus H_1$ leads to an accessible path from $\hat{\mathbf{0}}$ to $\hat{\mathbf{1}}$, the probability for the latter is bounded by

$$\mathbb{P}[X_{L,\alpha} > 0] \geq q_L\left[1 - (1 - \mathbb{P}[X_{L-2,\tilde{\alpha}} > 0])\left(1 - \frac{\mathbb{P}[X_{L-2,\tilde{\alpha}} > 0]}{2}\right)\right]. \tag{3.45}$$

It follows from equation (3.39) that

$$\lim_{L\to\infty} \mathbb{P}[X_{L,\alpha} > 0] \geq 1 - \left(1 - \frac{1}{\tilde{C}_\delta}\right)\left(1 - \frac{1}{2\tilde{C}_\delta}\right). \tag{3.46}$$

This procedure can be repeated multiple times. In each step, the limiting constant $A_n$ is replaced by another constant

$$A_{n+1} = 1 - (1 - A_n)(1 - A_n/2) \tag{3.47}$$

with $A_0 = 1/\tilde{C}_\delta$. The fixed points of this recurrence relation are given by 0 and 1. Furthermore, the resulting sequence is monotonically increasing as long as $0 < A_0 < 1$. Therefore, $A_n \to 1$ for $n \to \infty$ and as a consequence

$$\lim_{L\to\infty} \mathbb{P}[X_{L,\alpha} > 0] = 1\,.$$

Now one can finally show the result (3.22). Let $\alpha = (\log L - \epsilon_L)/L$ with $\epsilon_L \to \infty$ and $\epsilon_L/\log L \to 0$. Then

$$\begin{aligned}\mathbb{P}[X_L > 0]\,\frac{L}{\log L} &\geq \mathbb{P}\big[X_L > 0 \mid w(\hat{\mathbf{0}}) < \alpha\big]\,\mathbb{P}\big[w(\hat{\mathbf{0}}) < \alpha\big]\,\frac{L}{\log L} \\ &\geq \mathbb{P}[X_{L,\alpha} > 0]\,\frac{L\,\alpha}{\log L} = \mathbb{P}[X_{L,\alpha} > 0]\left(1 - \frac{\epsilon_L}{\log L}\right) \xrightarrow{L\to\infty} 1\,,\end{aligned} \tag{3.48}$$

i.e., the lower limit for the asymptotic behavior is given by $\log L/L$. Together with equation (3.25), the said result follows.

### 3.2.4. Further Results

If $w(\hat{\mathbf{0}}) = \alpha = x/L$ with $x$ fixed, $X_{L,\alpha}/L$ converges in law to $\exp(-x)$ times the product of two independent random variables that are distributed according to a standard exponential distribution [52].

## 3.3. HoC Model on the Undirected Hypercube

Again, it is assumed that the global maximum is located at $\hat{\mathbf{1}}$ while the starting sequence of all paths under consideration is $\hat{\mathbf{0}}$. These paths do not need to be shortest on the undirected hypercube, i.e., "backsteps" are allowed. Furthermore, it is assumed that $w(\hat{\mathbf{0}}) = \alpha$, like in the case of the directed hypercube. The number of paths will be denoted by $X_\alpha$, where the explicit $L$-dependence is omitted as it will not be needed. In case $w(\hat{\mathbf{0}})$ is random, the number of paths is simply denoted by $X$. Obviously, an accessible path through the directed hypercube will also be a valid path through the undirected hypercube and hence $X_\alpha^{(\mathrm{ud})} > X_{L,\alpha}^{(\mathrm{d})}$, where the superscript "ud" and "d" is meant as an abbreviation for undirected and directed, respectively. Therefore, results on lower limits for the directed case are inherited by the undirected hypercube.

Since the study of accessible paths on the undirected hypercube is much more complicated, only the result for the first moment $\mathbb{E}[X_\alpha]$, which was obtained by Berestycki et al. [22], will be presented here. It states that

$$\lim_{L\to\infty} \mathbb{E}[X_\alpha]^{1/L} = \sinh(1-\alpha)\,, \tag{3.49}$$

which already implies for $\alpha > \alpha^* = 1 - \sinh^{-1}(1)$ that

$$\lim_{L\to\infty} \mathbb{P}[X_\alpha > 0] = 0\,. \tag{3.50}$$

This result was later completed independently by Martinsson [46] and Li [55] to

$$\lim_{L\to\infty} \mathbb{P}[X_\alpha > 0] = \begin{cases} 0 & \text{for } \alpha > \alpha^*\,, \\ 1 & \text{for } \alpha < \alpha^*\,, \end{cases} \tag{3.51}$$

and hence

$$\lim_{L\to\infty} \mathbb{P}[X > 0] = \alpha^* \approx 0.1186 \tag{3.52}$$

for unconstrained fitness of $\hat{\mathbf{0}}$.

For the proof of (3.49) one needs estimates of the number $a_{L,p}$ of paths that include $p$ backsteps. Each backstep must be compensated by an additional step in the direction of $\hat{\mathbf{1}}$ and hence such a path consists of $L + 1 + 2p$ vertices. Due to the constraints on $\hat{\mathbf{0}}$ and $\hat{\mathbf{1}}$, the first moment can be written as

$$\mathbb{E}[X_\alpha] = \sum_{p=0}^{\infty} \frac{a_{L,p}\,(1-\alpha)^{L+2p-1}}{(L+2p-1)!}\,. \tag{3.53}$$

Note that $a_{L,p} = 0$ for sufficiently large $p$ as the length of a self-avoiding path cannot exceed the total number $2^L$ of vertices on the hypercube. The idea is now to find upper and lower bounds for $a_{L,p}$ that maintain the correct asymptotic behavior of $\mathbb{E}[X_\alpha]$.

### 3.3.1. Upper bound for $a_{L,p}$

A rather simple upper bound arises by neglecting the constraint that the paths is self-avoiding. Let $B_{L,p}$ be the number of those paths, i.e., arbitrary paths that connect $\hat{\mathbf{0}}$ to $\hat{\mathbf{1}}$ and contain $L + 1 + 2p$ vertices. By definition, $B_{L,p}$ is larger than $a_{L,p}$.

The ansatz to evaluate this number is a recurrence relation. Assume a path through the $L$-dimensional hypercube that contains $L + 1 + 2p - 2q$ vertices, i.e., a path that contributes to $B_{L,p-q}$. If an $(L + 1)$-th locus is added and flipped at $2q + 1$ arbitrary positions along the path, a valid path containing $L+2+2p$ vertices emerges, i.e., a path that contributes to $B_{L+1,p}$. Any valid path can be constructed in that way for some integer $q \in [0, p]$ and hence

$$B_{L+1,p} = \sum_{q=0}^{p} \binom{L+1+2p}{2q+1} B_{L,p-q}\,, \tag{3.54}$$

with $B_{1,p} = 1$. This relation can now be applied to the generating function

$$G_L(z) = \sum_{p=0}^{\infty} \frac{B_{L,p}}{(L+2p)!} z^{L+2p}\,, \tag{3.55}$$

which reads

$$\begin{aligned} G_{L+1}(z) &= \sum_{p=0}^{\infty} \frac{B_{L+1,p}}{(L+2p+1)!} z^{L+2p+1} \\ &= \sum_{p=0}^{\infty} \sum_{q=0}^{p} \binom{L+1+2p}{2q+1} \frac{B_{L,p-q}}{(L+2p+1)!}\, z^{L+2p+1} \\ &= \sum_{q=0}^{\infty} \sum_{p=q}^{\infty} \frac{B_{L,p-q}}{(2q+1)!\,[L+2(p-q)]!}\, z^{L+2(p-q)}\, z^{2q+1} \\ &= \sum_{q=0}^{\infty} \frac{z^{2q+1}}{(2q+1)!} \sum_{p=0}^{\infty} \frac{B_{L,p}}{(L+2p)!} z^{L+2p} = \sinh(z)\, G_L(z)\,. \end{aligned} \tag{3.56}$$

Since $G_1(z) = \sinh(z)$, the solution of this recurrence relation is given by

$$G_L(z) = \sinh(z)^L\,. \tag{3.57}$$

Finally, according to equation (3.53) and (3.55), the first moment can be bounded from above by

$$\begin{aligned} \mathbb{E}[X_\alpha] &\leq \sum_{p=0}^{\infty} \frac{B_{L,p}\,(1-\alpha)^{L+2p-1}}{(L+2p-1)!} = \left.\frac{\mathrm{d}G_L(z)}{\mathrm{d}z}\right|_{z=1-\alpha} \\ &= L\,\sinh(1-\alpha)^L \operatorname{cotanh}(1-\alpha)\,. \end{aligned} \tag{3.58}$$

and hence

$$\lim_{L\to\infty} \mathbb{E}[X_\alpha]^{1/L} \leq \sinh(1-\alpha)\,. \tag{3.59}$$

### 3.3.2. Lower bound for $a_{L,p}$

As a lower bound one can use the subset of paths that contribute to $a_{L,p}$ and stay valid self-avoiding paths through the $(L-1)$-dimensional hypercube if the $L$-th locus is removed. Let $b_{L,p}$ be the number of these paths. For instance, the path `13121` is a valid path contributing to $a_{3,1}$ but not to $b_{3,1}$, because if `3` is removed, the path `1121` is not self-avoiding anymore.

Similar to $B_{L,p}$, one can estimate $b_{L,p}$ recursively. Suppose a path through the $L$-dimensional hypercube contributing to $b_{L,p-q}$. If an $(L+1)$-th locus is added that is flipped $2q+1$ times, a valid path contributing to $b_{L+1,p}$ emerges if and only if the locus $(L+1)$ is not flipped consecutively. In other words, from the $(L+1+2p)$ times that a

locus is flipped along the path, one can chose $2q+1$ times where the $(L+1)$-th locus is flipped without flipping it two times in a row. This gives $\binom{L+1+2p-2q}{2q+1}$ possibilities and hence the analogue to equation (3.54) reads

$$b_{L+1,p} = \sum_{q=0}^{p} \binom{L+1+2(p-q)}{2q+1} b_{L,p-q}\,, \tag{3.60}$$

where $b_{1,p} = \delta_{p,0}$. Since

$$\begin{aligned}\binom{L+1+2(p-q)}{2q+1} &= \binom{L+1+2p}{2q+1} \prod_{r=0}^{2q-1} \left[1 - \frac{2q+1}{L+1+2p-r}\right] \\ &\leq \binom{L+1+2p}{2q+1} \left[1 - \frac{2q+1}{L+2}\right]^{2q}\end{aligned}$$

for $q < p$, one can define $\tilde{b}_{L,p}$ which obeys the simpler recurrence relation

$$\tilde{b}_{L+1,p} = \sum_{q=0}^{p} \binom{L+1+2p}{2q+1} \left[1 - \frac{2q+1}{L+2}\right]^{2q} \tilde{b}_{L,p-q} \tag{3.61}$$

and fulfills that

$$\tilde{b}_{L,p} \leq b_{L,p} \leq a_{L,p}\,. \tag{3.62}$$

The recurrence relation for $\tilde{b}_{L,p}$ is again handed on to its generating function

$$g_L(z) = \sum_{p=0}^{\infty} \frac{\tilde{b}_{L,p}}{(L+2p)!}\, z^{L+2p}\,. \tag{3.63}$$

This yields

$$\begin{aligned} g_{L+1}(z) &= \sum_{p=0}^{\infty} \frac{\tilde{b}_{L+1,p}}{(L+2p+1)!}\, z^{L+2p+1} \\ &= \sum_{p=0}^{\infty} \sum_{q=0}^{p} \binom{L+1+2p}{2q+1} \left[1 - \frac{2q+1}{L+2}\right]^{2q} \frac{\tilde{b}_{L,p-q}}{(L+2p+1)!}\, z^{L+2p+1} \\ &= \sum_{q=0}^{\infty} \left[1 - \frac{2q+1}{L+2}\right]^{2q} \frac{z^{2q+1}}{(2q+1)!} \sum_{p=q}^{\infty} \frac{\tilde{b}_{L,p-q}}{[L+2(p-q)]!} z^{L+2(p-q)} \\ &= \sinh_{L+1}(z)\, g_L(z)\,, \end{aligned} \tag{3.64}$$

where

$$\sinh_k(z) = \sum_{q=0}^{\infty} \left[1 - \frac{2q+1}{k+1}\right]^{2q} \frac{z^{2q+1}}{(2q+1)!}\,. \tag{3.65}$$

In analogy to ordinary hyperbolic functions, one may also define

$$\cosh_k(z) = \frac{\mathrm{d}\sinh_k(z)}{\mathrm{d}z} = \sum_{q=0}^{\infty}\left[1 - \frac{2q+1}{k+1}\right]^{2q} \frac{z^{2q}}{(2q)!}\,. \tag{3.66}$$

Since $g_1(z) = \sinh_1(z)$, the solution of equation (3.64) is given by

$$g_L(z) = \prod_{k=1}^{L} \sinh_k(z)\,. \tag{3.67}$$

Again, one is interested in the derivative which yields

$$\mathbb{E}[X_\alpha] \geq \left.\frac{\mathrm{d}g_L(z)}{\mathrm{d}z}\right|_{z=1-\alpha} = g_L(1-\alpha) \sum_{k=1}^{L} \frac{\cosh_k(1-\alpha)}{\sinh_k(1-\alpha)}\,. \tag{3.68}$$

One still needs to compute the $L \to \infty$ behavior of the right-hand side of that inequality. By definition, the pseudo hyperbolic functions are bounded by $z \leq \sinh_k(z) \leq \sinh(z)$ and $1 \leq \cosh_k(z) \leq \cosh(z)$, respectively, and hence

$$L\,\frac{1}{\sinh z} \leq \sum_{k=1}^{L} \frac{\cosh_k(z)}{\sinh_k(z)} \leq L\,\frac{\cosh(z)}{z}\,,$$

which in turn implies according to the squeeze theorem that

$$\lim_{L\to\infty}\left(\sum_{k=1}^{L} \frac{\cosh_k(z)}{\sinh_k(z)}\right)^{1/L} = 1\,.$$

Furthermore, it follows from dominated convergence that

$$\lim_{k\to\infty} \sinh_k(z) = \sinh(z)\,,$$

and hence also the geometric Cesàro mean converges, i.e.,

$$g_L(z)^{1/L} = \left[\prod_{k=1}^{L} \sinh_k(z)\right]^{1/L} \xrightarrow{L\to\infty} \sinh(z)\,. \tag{3.69}$$

According to equation (3.68) this yields

$$\lim_{L\to\infty} \mathbb{E}[X_\alpha]^{1/L} \geq \sinh(1-\alpha)\,, \tag{3.70}$$

which implies together with equation (3.59) the result (3.49).

### 3.3.3. Further Results

On the undirected hypercube one can also ask what happens when the global optimum is not located at $\hat{\mathbf{1}}$ but at some arbitrary sequence $\sigma_{\max}$. If only shortest paths are allowed, this would be basically the same problem, but on a hypercube with lowered dimension $L' = d(\hat{\mathbf{0}}, \sigma_{\max})$. In the undirected case here, however, this is a non-trivial problem.

Let $X_{\alpha,y}$ be the number of accessible paths from $\hat{\mathbf{0}}$ to $\sigma_{\max}$, given that $w(\hat{\mathbf{0}}) = \alpha$ and $d(\hat{\mathbf{0}}, \sigma_{\max})/L = y$. It was shown in [22] that

$$\lim_{L\to\infty} \mathbb{E}[X_{\alpha,y}]^{1/L} = \sinh(1-\alpha)^y \, \cosh(1-\alpha)^{1-y} \,, \tag{3.71}$$

which reduces for $y = 1$ nicely to the case where $\sigma_{\max} = \hat{\mathbf{1}}$. Concerning the actual percolation probability, it was proven in [46, 55] that

$$\lim_{L\to\infty} \mathbb{P}[X_{\alpha,y} > 0] = \begin{cases} 0 & \text{for } \alpha > \alpha^* \,, \\ 1 & \text{for } \alpha < \alpha^* \,, \end{cases} \tag{3.72}$$

where $\alpha^*$ is the unique positive solution of

$$\sinh(1-\alpha)^y \, \cosh(1-\alpha)^{1-y} = 1 \tag{3.73}$$

with respect to $\alpha$. The case where $\sigma_{\max}$ is chosen at random corresponds to $y = 1/2$ [46].

## 3.4. RMF Model

As will be shown in this section, accessibility percolation on RMF landscapes is similar to ordinary site percolation which makes it actually simpler to study than on HoC landscapes.

### 3.4.1. Ordering Probability

To begin with, consider the probability $P_{\text{order}}$ that a given path of length $L$ parallel to the mean fitness gradient is accessible. To detach this problem from the full problem, define $\tilde{Y}_k = Y_k + ks$ for $1 \leq k \leq L$, where the $Y_k$ are i.i.d. random variables drawn from a distribution with cumulative distribution function $F$. The slope $s > 0$ has the same meaning as for the RMF model. In the literature, this model is known as the *linear drift model* [56] and was introduced in the context of record statistics. The relevant quantity for accessibility is the ordering probability

$$P_{\text{order}} = \mathbb{P}\Big[\tilde{Y}_1 < \tilde{Y}_2 < \cdots < \tilde{Y}_L\Big] \,. \tag{3.74}$$

It can be written formally as

$$P_{\text{order}} = \int_{-\infty}^{\infty} \mathrm{d}y_L \, f(y_L) \int_{-\infty}^{y_L+s} \mathrm{d}y_{L-1} \, f(y_{L-1}) \cdots \int_{-\infty}^{y_2+s} \mathrm{d}y_1 \, f(y_1) \,, \tag{3.75}$$

but only be expressed in closed form for few special cases. It was shown in [57] for a Gumbel distribution, i.e., $F(x) = \exp[-\exp(-x)]$, that

$$P_{\text{order}}(L) = a(s)^L \, b(L,s) \,, \tag{3.76}$$

where

$$a(s) = 1 - e^{-s} \qquad \text{and} \qquad b(L,s) = \prod_{k=1}^{L} [1 - \exp(-sk)]^{-1} \,.$$

Note that $b(L,s)$ converges to a positive constant $b_\infty(s)$ for $L \to \infty$ and hence the ordering probability decays asymptotically exponential. For general distributions, one can easily find an exponentially decaying lower limit. If all the random parts $Y_i$ are in the same interval of length $s$, the $\tilde{Y}_i$ will be ordered due to the drift term. Therefore,

$$P_{\text{order}} \geq \left( \max_x \int_x^{x+s} f(y) \, \mathrm{d}y \right)^L . \tag{3.77}$$

The similarity to ordinary site percolation arises from the fact that there the probability for a path to be open also decays exponentially as $q^L$ if sites are independently open with probability $q$.

### 3.4.2. RMF Model on Regular Trees

The RMF model can be simply defined on trees with the root $\hat{\mathbf{0}}$ playing the role of the reference sequence on the hypercube, i.e., for an arbitrary node $\sigma$, $w(\sigma)$ is distributed according to the PDF $f(x - d(\sigma, \hat{\mathbf{0}})s)$. It will be assumed in the following that $f$ is the PDF of a Gumbel distribution such that equation (3.76) is applicable. Note that $b(L,s)$ is monotonically increasing in $L$ and hence

$$1 \leq b(L,s) \leq b_\infty(s) \,. \tag{3.78}$$

The first moment of the number of accessible paths $X$ can then be estimated by

$$\mathbb{E}[X] = n^L \, P_{\text{order}}(L+1) \leq [n \, a(s)]^L a(s) \, b_\infty(s) \,. \tag{3.79}$$

Depending on whether $n < a(s)$ or $n > a(s)$, the limiting value of the first moment is given by zero or infinity, respectively, and hence there is a threshold $s_{\text{crit}} = \log(n) - \log(n-1)$ where this transition happens. For $s < s_{\text{crit}}$, the probability of having accessible paths vanishes since

$$\mathbb{P}[X > 0] \leq \mathbb{E}[X] \leq [n \, a(s)]^L a(s) \, b_\infty(s) \xrightarrow{L \to \infty} 0. \tag{3.80}$$

Now let $s > s_{\text{crit}}$. In order to apply inequality (3.10), one needs to estimate the second moment of $X$. A simple upper bound arises from the fact that if two paths with $k+1$ common vertices are accessible, the fitness of the common vertices have to be ordered as well as the fitness along the two subpaths consisting of $L - k$ separated vertices.

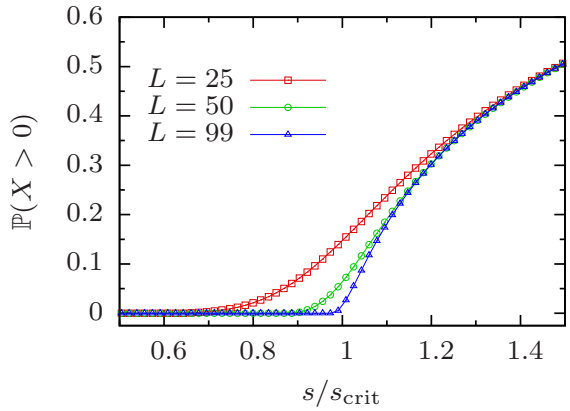



**Figure 3.5.** Probability $\mathbb{P}[X > 0]$ to have accessible paths on an $n = 4$ tree in dependence on the slope $s$ of the RMF model with Gumbel distributed random part. The critical slope is given by $s_{\text{crit}} = \log(4/3)$. Symbols correspond to the numerical evaluation of equation (3.2), lines are for visual guidance.

Therefore, the probability $p_k$ that both paths are accessible can be bounded from above by

$$p_k \leq P_{\text{order}}(k+1)\, P_{\text{order}}(L-k)^2 \tag{3.81}$$

and hence, analogously to equation (3.8),

$$\begin{aligned}\mathbb{E}\big[X^2\big] &= \sum_{i=1}^{n^L}\sum_{j=1}^{n^L}\mathbb{E}[\theta_i\,\theta_k] = \sum_{k=0}^{L} p_k\, m_k \leq \sum_{k=0}^{L} a(s)^{2L-k+1} b_\infty(s)^3 n^{2L-k} \\ &= \frac{b_\infty(s)^3\, a(s)\,[a(s)n]^L\big[(a(s)n)^{L+1}-1\big]}{a(s)\,n-1}\,.\end{aligned} \tag{3.82}$$

Applying inequality (3.10) finally yields

$$\begin{aligned}\mathbb{P}[X>0] \geq \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} &\geq \frac{[n\,a(s)]^{2L} a(s)^2\,[a(s)\,n-1]}{b_\infty(s)^3\,a(s)\,[a(s)n]^L[(a(s)n)^{L+1}-1]} \\ &\xrightarrow{L\to\infty} \frac{a(s)\,n-1}{b_\infty(s)^3\,n} > 0\,.\end{aligned} \tag{3.83}$$

Combined with equation (3.80), there is a transition at $s_{\text{crit}}$ where

$$\lim_{L\to\infty}\mathbb{P}[X>0]\begin{cases} = 0 & \text{for } s < s_{\text{crit}}\,, \\ > 0 & \text{for } s > s_{\text{crit}}\,. \end{cases} \tag{3.84}$$

This result can also be obtained numerically by the evaluation of equation (3.2) as shown in figure 3.5.

Technically, this method works also for arbitrary distributions, i.e., there is a percolation threshold defined implicitly by $a(s_{\text{crit}}) = n^{-1}$, as long as the ordering probability can be written like in equation (3.76) with $b(L, s)$ that is bounded for $L \to \infty$, even though I am not aware of any other examples. Nevertheless, as can be seen from equation (3.77), one can always find a sufficiently large value of $s$ such that

$$a(s) \geq \max_x \int_x^{x+s} f(y)\,\mathrm{d}y > n^{-1}$$

and in turn $\lim_{L\to\infty} \mathbb{P}[X > 0] > 0$. Furthermore, for $s = 0$ one has obviously $\mathbb{P}[X > 0] \to 0$ and hence there is always a percolation threshold $s_{\text{crit}} \geq 0$.

### 3.4.3. RMF Model on Hypercubes

In this section, the strategy of Hegarty and Martinsson in [18] will be used again in order to show that there is always an accessible path on the RMF landscape for $L \to \infty$ and $s > 0$. In fact, this does also hold true if $s = s_L$ is a sequence with $s_L \to 0$ as long as $L\,s_L \to \infty$. The ansatz is to show first that there is always a path through the hypercube in case of ordinary site percolation which can then be extended to accessibility percolation on the RMF landscape. In the following, it will be assumed without loss of generality that the reference sequence of the RMF model is given by the all-zero sequence, i.e., $\tilde{\sigma} = \hat{\mathbf{0}}$. Like before, accessible paths are required to connect $\hat{\mathbf{0}}$ with $\hat{\mathbf{1}}$, but now the latter sequence is not necessarily the global maximum.

Consider ordinary site percolation for now. As already discussed, this type of percolation means that vertices of the hypercube are independently removed with probability $(1 - q)$. A path is called *open* if all vertices along the paths are present. Let $Z_q$ be the number of open paths from $\hat{\mathbf{0}}$ to $\hat{\mathbf{1}}$, given that both $\hat{\mathbf{0}}$ and $\hat{\mathbf{1}}$ are present. The expected value of $Z_q$ is given by

$$\mathbb{E}[Z_q] = L!\,q^{L-1}\,, \tag{3.85}$$

which goes to infinity for any $q > 0$. In order to compute the second moment, let $\zeta_i$ be the indicator variable that the $i$-th path is open. Obviously, a given path is open with probability $q^{L-1}$, a pair of paths that shares $k-1$ interior vertices is open if all $2L-k-1$ vertices are present, i.e., with probability $q^{2L-k-1}$. In analogy to equation (3.34) one gets

$$\begin{aligned}\mathbb{E}\big[Z_q^2\big] &= \sum_{i=1}^{2^L}\sum_{j=1}^{2^L} \mathbb{E}[\zeta_i\zeta_j] = \mathbb{E}[Z_q] + L!\sum_{k=1}^{L} T(L,k)\,q^{2L-k-1} \\ &= \mathbb{E}[Z_q] + \mathbb{E}[Z_q]^2 \left( \frac{T(L,1)}{L!} + \underbrace{\sum_{k=2}^{(1-\delta)L} \frac{T(L,k)}{L!\,q^{k-1}}}_{S_1} + \underbrace{\sum_{k=0}^{\delta L} \frac{T(L,L-k)}{L!\,q^{L-k-1}}}_{S_2} \right). \end{aligned} \tag{3.86}$$

As can be seen from equation (2.1), the ratio $T(L,1)/L!$ converges to 1. $S_1$ can be estimated with equation (3.32) and Stirling's formula that yields

$$\begin{aligned} S_1 &= \sum_{k=2}^{(1-\delta)L} \frac{T(L,k)}{L!\,q^{k-1}} \leq C_\delta \sum_{k=2}^{(1-\delta)L} \frac{k\,(L-k+1)!}{L!\,q^{k-1}} \\ &\leq \frac{C_\delta}{\sqrt{2\pi}} \sum_{k=2}^{(1-\delta)L} \frac{k\,e^k\,(L-k+1)^{L-k+3/2}}{L^{L+1/2}\,q^{k-1}} \\ &\leq \frac{C_\delta}{\sqrt{2\pi}} \sum_{k=2}^{\infty} k\left(\frac{e}{q\,L}\right)^{k-1} = \frac{C_\delta}{\sqrt{2\pi}}\left[\left(1-\frac{e}{q\,L}\right)^{-2} - 1\right] \xrightarrow{L\to\infty} 0\,. \end{aligned} \tag{3.87}$$

For the second part $S_2$, equation (3.33) will be used which results in

$$\begin{aligned} S_2 &= \frac{1}{\mathbb{E}[Z_q]} \sum_{k=0}^{\delta L} T(L,L-k) q^k \leq \frac{c}{\mathbb{E}[Z_q]} \sum_{k=0}^{\delta L} (k+1)\left(q\,\frac{L+2k}{5}\right)^k \\ &\leq \frac{c}{\mathbb{E}[Z_q]} \sum_{k=0}^{\delta L} [(1+2\delta)\,q\,L]^k = \frac{c}{\mathbb{E}[Z_q]}\,\frac{[(1+2\delta)\,q\,L]^{\delta L+1}-1}{[(1+2\delta)\,q\,L]-1} \xrightarrow{L\to\infty} 0\,. \end{aligned} \tag{3.88}$$

Finally, one can use equation (3.10) again to obtain

$$\mathbb{P}[Z_q > 0] \geq \frac{\mathbb{E}[Z_q]^2}{\mathbb{E}\left[Z_q^2\right]} \geq \left(\frac{1}{\mathbb{E}[Z_q]} + \frac{T(L,1)}{L!} + S_1 + S_2\right)^{-1} \xrightarrow{L\to\infty} 1\,. \tag{3.89}$$

Note that this result is also valid for $q \to 0$ as long as $q\,L \to \infty$, since $S_1$, $S_2$ and $1/\mathbb{E}[Z_q]$ also converge to zero in that case.

Now consider the RMF landscape. The accessibility percolation problem will be mapped onto the ordinary site percolation problem. Let $\tilde{w}(\sigma) = w(\sigma) - s\,d(\sigma,\hat{\mathbf{0}})$ be the random part of fitness. Furthermore, $\tilde{w}(\hat{\mathbf{0}})$ and $\tilde{w}(\hat{\mathbf{1}})$ are assumed to be fixed, but may be arbitrary. In case $\tilde{w}(\hat{\mathbf{0}}) - \tilde{w}(\hat{\mathbf{1}}) < s$, define each vertex $\sigma$ as open for which

$$\tilde{w}(\sigma) \in J_s = \left[\frac{\tilde{w}(\hat{\mathbf{0}}) + \tilde{w}(\hat{\mathbf{1}})}{2} - s/2,\, \frac{\tilde{w}(\hat{\mathbf{0}}) + \tilde{w}(\hat{\mathbf{1}})}{2} + s/2\right]. \tag{3.90}$$

If there is an open path, there will be also an accessible path since the random parts of all fitness values lie in the same interval of size $s$ and a slope of $s$ will "lift" the random variables such that they are ordered. Therefore,

$$\mathbb{P}[X > 0] \geq \mathbb{P}[Z_q > 0] \xrightarrow{L\to\infty} 1\,, \tag{3.91}$$

with

$$q = \mathbb{P}[\tilde{w}(\sigma) \in J_s]\,. \tag{3.92}$$

In case $\Delta\tilde{w} = \tilde{w}(\hat{\mathbf{0}}) - \tilde{w}(\hat{\mathbf{1}}) > s$, one needs to bridge the gap between $w(\hat{\mathbf{0}})$ and $w(\hat{\mathbf{1}})$. To do so, define a vertex $\sigma$ with $d(\sigma, \hat{\mathbf{0}}) = d$ as open if

$$\tilde{w}(\sigma) \in J_s(d) = \left[\tilde{w}(\hat{\mathbf{1}}) + \frac{d-1}{L}\Delta\tilde{w}, \tilde{w}(\hat{\mathbf{1}}) + \frac{d-1}{L}\Delta\tilde{w} + \frac{s}{2}\right] \tag{3.93}$$

for $1 \leq d \leq L-1$. As long as $s/2 > \Delta w/L$, which is ensured for $s\,L \to \infty$, this family of intervals will cover the whole interval $[\tilde{w}(\hat{\mathbf{1}}), \tilde{w}(\hat{\mathbf{0}})]$. If so, an open path will also be accessible since by construction of $J_s(d)$ it follows from $\tilde{w}(\sigma) \in J_s(d)$ and $\tilde{w}(\tau) \in J_s(d+1)$ that $\tilde{w}(\tau) + s > \tilde{w}(\sigma)$. Similar to the previous case, one has now

$$\mathbb{P}[X > 0] \geq \mathbb{P}[s/2 > \Delta w/L]\,\mathbb{P}[Z_q > 0] \xrightarrow{L\to\infty} 1\,, \tag{3.94}$$

where

$$q = \min_d \mathbb{P}[\tilde{w}(\sigma) \in J_s(d)]\,. \tag{3.95}$$

This implies that accessibility paths can be found for all $\tilde{w}(\hat{\mathbf{0}})$ and $\tilde{w}(\hat{\mathbf{1}})$ as $L \to \infty$. Furthermore, $q$ defined in equations (3.92) and (3.95) will become proportional to $s$ for $s \to 0$ since it is given by the integral over a (probability density) function in a small interval of size $s$ and $s/2$, respectively. Therefore, $s \to 0$ and $s\,L \to \infty$ implies also $q\,L \to \infty$ which in turn implies $\mathbb{P}[X > 0] \geq \mathbb{P}[Z_q > 0] \to 1$ in that case.

## 3.5. NK Model

### 3.5.1. Block Model

Like in case of local maxima, the easiest analytical access to accessible paths through the NK landscape is for blockwise interactions as results that were already obtained for the HoC model can be used again. Given that $\hat{\mathbf{1}}$ is the global maximum, every single block must have its maximal fitness value at that point, too. An accessible path from $\hat{\mathbf{0}}$ to $\hat{\mathbf{1}}$ exists if and only if there is an accessible path through each sublandscape defined by the blocks. As already mentioned, the blocks are independent and hence the probability to have an accessible path through the block model is given by

$$\mathbb{P}[X_{L,K} > 0] = \mathbb{P}[X_K > 0]^{\frac{L}{K}} \approx \left(\frac{\log K}{K}\right)^{\frac{L}{K}}\,, \tag{3.96}$$

where $X_{L,K}$ is the number of paths through the full NK landscape and $X_K$ is the number of paths in a HoC landscape of size $K$. According to equation (3.22), the approximation is asymptotically exact for $K \to \infty$.

Paths through the block model have a unique property: Suppose you have found an accessible path in the full landscape. This corresponds to the flipping of loci from 0 to 1 in a specific order and each flip takes place in a certain block. Any other order will also lead to a valid accessible path as long as the order within each block remains the

same. For instance, assume the path `1234` is accessible in a block model with $L = 4$ and $K = 2$. Then the paths `1324`, `1342`, `3124`, `3142` and `3412` are accessible as well since locus 1 is flipped before locus 2 in the first block and locus 3 is flipped before locus 4 in the second block. Thus the resulting subpaths through the blocks are all equivalent. In general, one can construct $L!/K!^{L/K}$ paths from each set of valid subpaths and hence

$$X_{L,K} = \frac{L!}{K!^{L/K}} \prod_{i=1}^{L/K} X_K^{(i)} \,, \tag{3.97}$$

where $X_K^{(i)}$ is the number of accessible paths through the $i$-th block. Since the $X_K^{(i)}$ are independent and $\mathbb{E}\left[X_K^{(i)}\right] = 1$ for all $i$, one immediately finds

$$\mathbb{E}[X_{L,K}] = \frac{L!}{K!^{L/K}} \,. \tag{3.98}$$

In the same manner, also higher moments or even the full distribution can be calculated [40] from equation (3.97). The interesting thing about the distribution is that the number of paths is always an integer multiple of its mean value $\mathbb{E}[X_{L,K}]$ which diverges for $L \to \infty$ while the probability $\mathbb{P}[X_{L,K} > 0]$ decays exponentially.

### 3.5.2. Adjacent and Random Interactions

For a long time, results on the accessibility of $\sigma_{\max}$ were only obtained numerically [28, 40, 58] which restricts the genome length $L$ to rather small values. As shown in figure 3.6, the behavior of $\mathbb{P}[X > 0]$ seems qualitatively different for the three interaction types. For $K = 2$, this quantity still decreases monotonically with $L$ and all interaction schemes look similar, but notable differences arise for $K \geq 3$. In case of $K = 3$, only block and adjacent interactions lead to a decreasing accessibility, but $\mathbb{P}[X > 0]$ seems to be roughly constant for random interactions. If $K$ is increased further, the decline of adjacent interactions vanishes and random interactions even allow for increasing accessibility with $L$. Generally, the accessibility increases if one switches from block to adjacent and then to random interactions, which is consistent with the finding from section 2.4 that the number $N_{\max}$ of local maxima decreases in the same manner. Note, however, that these results on accessibility are only valid for small values of $L$ and $K$. The order of $\mathbb{P}[X > 0]$ with respect to the interaction scheme remains presumably the same for large $L$. Nevertheless, these schemes have one thing in common: The accessibility goes to zero for $L \to \infty$. The reason is similar to that for the inaccessibility in the block model, namely that the genotype can be divided into regions that have an independent chance to be inaccessible. For $L \to \infty$, the number of those regions grows indefinitely and hence the probability that at least one region is inaccessible goes to one.

In case of adjacent interactions, this phenomenon is still relatively simple to explain. Based on the idea in [59], suppose a sub-landscape of size 2 that emerges if one keeps all loci but $i$ and $i+1$ constant. By construction of the interaction sets, $V_{i+1-K}$ contains $i$, $V_{i+1}$ contains $i+1$ and $V_j$ contains both loci for $i+1-K < j < i+1$. A locus that is

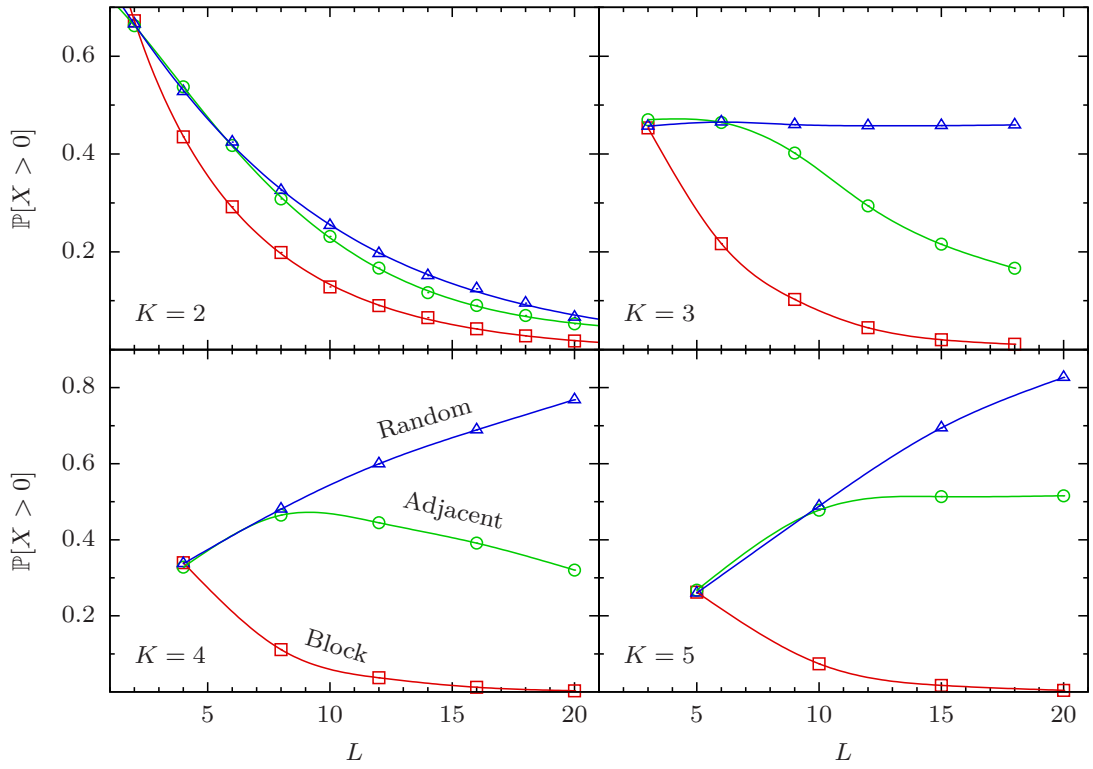



**Figure 3.6.** Numerically obtained probabilities $\mathbb{P}[X > 0]$ to have accessible paths in an NK landscape for various values of $K$. Different symbols correspond to different interaction patterns and are defined in the $K = 4$ panel. Lines are for visual guidance.

not contained in any of these sets does not affect the two-locus sub-landscape, or to put it the other way round, a locus $l$ influences the sub-landscape if and only if

$$l \in \{i - K + 1, i - K + 2, \ldots, i - 1, i + 2, i + 3, \ldots, i + K\} .$$

The set contains $2K - 2$ elements, i.e., the sub-landscape plus all loci that influence it are $2K$ loci in total. Thus one can construct up to $\lfloor L/2K \rfloor$ of them without overlapping. If the sub-landscape consisting of $i$ and $i+1$ has a local maximum at $(i, i+1) = (0, 0)$ for all configurations of loci influencing it, the whole landscape will be inaccessible as well since neither the $i$-th nor the $(i+1)$-th locus can be flipped from 0 to 1. The probability $P_{\text{sub}}$ for this to happen might be small but is non-zero and independent of $L$ which yields

$$\mathbb{P}[X > 0] < (1 - P_{\text{sub}})^{\lfloor L/2K \rfloor} \xrightarrow{L \to \infty} 0 . \tag{3.99}$$

The situation for random interactions is unfortunately much more complicated. One can still divide the genotype into independent regions: Each interaction set has roughly a probability of $2K/L$ to contain at least either of the two loci of a pair $(i, j)$ and hence on average there are $2K$ sets involved, each of which contains $K$ loci. Therefore, the

number of loci involved in a sub-landscape of size two is expected to be of order $K^2$ and hence the expected number of independent regions is of order $L/K^2$, which also diverges linearly. The problem is that the probability $P_{\text{sub}}$ depends on the numbers of interaction sets that contain $i$, $j$ and both, respectively. That a set contains both loci is given by the probability $K\,(K-1)/(L-1)$ while the probability that it only contains one of the loci is $2\,K\,(L-K)/(L-1)$. The latter is substantially larger for $L \gg K$ and the resulting sub-landscape will have a much bigger additive part that reduces the number of maxima. For this reason, $P_{\text{sub}}$ is $L$-dependent and goes to zero for $L \to \infty$ which necessitates for a more careful analysis.

A rigorous proof for the inaccessibility of the NK model with random interactions was given in [59] that also includes a more general class of interaction schemes. The proof is based on the occurrence of *global reciprocal sign epistasis*, i.e., reciprocal sign epistasis between two loci $i$ and $j$ for all genetic backgrounds. If such a motif exists in the landscape, accessible paths cannot exist between antipodal sequences due to the impossibility to flip both $i$ and $j$, independent of the global optimums position. The probability that no pair of loci has this property was shown to decay exponentially and hence also $\mathbb{P}[X > 0]$ converges to zero at least exponentially.

It is still an open question how the percolation probability behaves if $K$ is scaled with $L$ as $L \to \infty$. If the ratio $L/K^\alpha$ is kept constant for some $\alpha > 0$, the number of independent regions, which is the actual cause of the inaccessibility, will grow more slowly and the size of the regions will become larger. Therefore, the arguments given above do not work anymore. Additionally, the NK landscape is still less rugged compared to a HoC landscape in terms of local maxima. This might be a hint that there exist a scaling $K = K(L)$ such that $\mathbb{P}[X > 0] \to 1$ for adjacent and random interactions.

# 4. Adaptive Walks

An adaptive walk (AW) is an idealized and rather simple evolutionary process. Instead of treating a population as a set of individuals, it is treated as a single entity that moves through the fitness landscape. Formally, adaptive walks are Markov processes. The state of the system is fully determined by the genotype $\sigma$ carried by the population and its dynamics by the transition probability $p(\sigma \to \tau)$ to "walk" to a new genotype $\tau$. This probability is zero if $d(\sigma, \tau) \neq 1$ or if $w(\tau) < w(\sigma)$ and thus adaptive walks terminate when a local maximum $\sigma_{\max}$ is reached. Therefore, paths taken by an adaptive walker are accessible in the same sense as defined in chapter 3. The main quantities of interest will be the walk length $\ell$, i.e., the number of steps that were taken until the walk terminates, and the walk height $h = w(\sigma_{\max})$, i.e., the fitness of the final genotype that the walker has reached.

## The Strong Selection and Weak Mutation Regime

Despite their simplicity, adaptive walks arise as the limiting behavior in the so-called *strong selection and weak mutation* (SSWM) regime of more general and realistic dynamics like the Wright-Fisher dynamics [60–63]. *Strong selection* means that $N|\Delta w| \gg 1$, where $\Delta w$ is the fitness advantage of a mutant. It causes that a mutant with $\Delta w > 0$ has a chance to fixate rapidly while mutants with $\Delta w < 0$ will die out quickly. Obviously, $\Delta w = w(\sigma) - w(\tau)$ is in general different for any two genotypes $\sigma$ and $\tau$ in a high-dimensional fitness landscape and can only be partially controlled via the fitness distribution, but $N$ can nevertheless be arbitrarily large. *Weak mutation* means that $N\mu \ll 1$, i.e., mutations are rare such that the timescale between the emergence of two mutants is much smaller then the typical fixation time. The occurrence of a double mutation in a single individual as well as the presence of multiple mutants at once are then very unlikely.

The dynamics in the SSWM limit looks like the following: The whole population is monomorphic and carries the genotype $\sigma$ until a mutant carrying genotype $\tau$ with $d(\sigma, \tau) = 1$ arises. If $w(\tau) < w(\sigma)$, the mutant and its offspring will die out in the long run. If $w(\tau) > w(\sigma)$, the genotype $\tau$ has a finite probability to fixate, but might vanish as well. This can happen multiple times before a fixation event occurs, but fixation will happen eventually as long as $\sigma$ is not a local maximum. Since fixation means that the whole population carries the new genotype $\tau$, this process starts from anew. In an adaptive walk, all of these things that may happen between fixation events are simplified into one single adaptive step defined by the transition probability $p(\sigma \to \tau)$.

### Types of Adaptive Walks

Depending on the details of how the SSWM limit is obtained, there are different types of adaptive walks regarding the transition probabilities $p(\sigma \to \tau)$. The most common types are defined in the following. Note that they all have in common that $p(\sigma \to \tau) = 0$ if either $w(\tau) < w(\sigma)$ or $d(\sigma, \tau) \neq 1$. For the sake of simplicity, this will not be stated explicitly in the formulas for the transition probabilities. Furthermore, it is useful to define

$$\mathcal{U}_\sigma^+ = \{\tau \in \mathcal{U}_\sigma \mid w(\tau) > w(\sigma)\}\,,$$

which is the set of fitter neighbors of $\sigma$.

**Random Adaptive Walk.** If one assumes that the first mutant with fitness advantage will fixate, all fitter genotypes have an equal chance to become prevalent. This leads to the random AW [24, 64–66] with transition probability

$$p(\sigma \to \tau) = \frac{1}{|\mathcal{U}_\sigma^+|}\,. \tag{4.1}$$

**Natural Adaptive Walk.** The most realistic version, as the name suggests, is the natural adaptive walk. It is assumed that mutants fixate with probability depending on their fitness advantage according to Kimura's formula [67, 68]. This leads to the transition probability

$$p(\sigma \to \tau) = \frac{w(\tau) - w(\sigma)}{\sum_{\sigma' \in \mathcal{U}_\sigma^+} w(\sigma') - w(\sigma)}\,. \tag{4.2}$$

Note that, in contrast to the other walk types defined here, the dynamics is influenced by the underlying fitness distribution.

**Greedy Adaptive Walk.** The greedy walk corresponds to a situation where fixation is only possible for the fittest genotype of the neighborhood, i.e., the transition probability is given by

$$p(\sigma \to \tau) = \begin{cases} 1 & \text{if } \tau = \max(\mathcal{U}_\sigma^+), \\ 0 & \text{else.} \end{cases} \tag{4.3}$$

This walk type can also be found as *gradient walk* in the literature [42, 69, 70]. In contrast to the random and natural AW, the greedy AW is deterministic on a given realization of the fitness landscape.

**Reluctant Adaptive Walk.** Basically the opposite of the greedy walk. In each step, the walker goes to the genotype with lowest available fitness that is still larger than the current one. Accordingly, the transition probability is given by

$$p(\sigma \to \tau) = \begin{cases} 1 & \text{if } \tau = \min(\mathcal{U}_\sigma^+), \\ 0 & \text{else.} \end{cases} \tag{4.4}$$

Like the greedy walk, the reluctant dynamics is deterministic. Note that this walk type does not seem to have a biological interpretation and is therefore rarely used in the biological literature, but it occurs sometimes in the context of spin glasses and optimization [71–74]. It should be thought of as a tool to analyze the structure of fitness landscapes rather than a realistic approximation to the actual dynamics of a population.

It is also worth mentioning, that all of these walk types can be defined by the transition probability

$$p(\sigma \to \tau) = \frac{[w(\tau) - w(\sigma)]^g}{\sum_{\sigma' \in \mathcal{U}_\sigma^+} [w(\sigma') - w(\sigma)]^g}\,, \tag{4.5}$$

where the "greed" $g \in \mathbb{R}$ is a parameter. For $g \to \infty$, the greedy walk is obtained, the natural walk for $g = 1$, the random AW for $g = 0$ and the reluctant walk for $g \to -\infty$.

## 4.1. Tree Approximation on the HoC Landscape

During each step of an adaptive walk, the walker "chooses" the next genotype from the $L$ neighbors of its current genotype. It might happen that certain genotypes occur several times during the walk, but this becomes very unlikely for large $L$. Therefore, it is justified to approximate the walk on a tree. This means for practical purposes that the state of the walker is simply determined by its current fitness rather than its genotype. In each step, $L$ new random numbers are drawn which represent the fitness values of neighboring genotypes. One of these numbers is then chosen according to the transition probability of the respective walk type.

Rather than having transition probabilities $p(\sigma \to \tau)$ between genotypes one needs the transition matrix $\gamma(y \to x)$ between fitness values, i.e., the probability density to have fitness $x$ in the next step given that the current fitness is $y$. Since a step is only possible if the state with fitness $y$ is not a local maximum, integrating over $x$ must yield

$$\int_y^\infty \gamma(y \to x)\,\mathrm{d}x = 1 - F(y)^L\,, \tag{4.6}$$

where $F$ is the cumulative fitness distribution function, i.e., the right-hand side is the probability that there is at least one allowed step.

Define $P_\ell(x)$ as the joint probability (density) that the walk has fitness $x$ in the $\ell$-th step and that the walk lasts at least $\ell$ steps. Then one can make use of the tree structure to construct a recurrence relation for $P_\ell(x)$: An adaptive walk has fitness $x$ in the $(\ell+1)$-th step if it had some fitness $y < x$ in the $\ell$-th step and then transitioned to $x$, i.e.,

$$P_{\ell+1}(x) = \int_{-\infty}^x P_\ell(y)\,\gamma(y \to x)\,\mathrm{d}y\,, \tag{4.7}$$

where $P_0(x) = f(x)$ is the probability density of fitness values. This ansatz was first introduced by Flyvbjerg and Lautrup who used it to study the random adaptive walk [66].

If the solution of (4.7) is known, one can extract the quantities of interest from it. For instance, the marginal probability $P_\ell$ that the walk lasts at least $\ell$ steps can be obtained via

$$\begin{aligned} P_{\ell+1} &= \int_{-\infty}^{\infty} P_{\ell+1}(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{x} \mathrm{d}y\, P_\ell(y)\,\gamma(y \to x) \\ &= \int_{-\infty}^{\infty} \mathrm{d}y\, P_\ell(y) \int_{y}^{\infty} \mathrm{d}x\,\gamma(y \to x) = \int_{-\infty}^{\infty} P_\ell(y)\left(1 - F(y)^L\right)\mathrm{d}y\,, \end{aligned}$$

where equation (4.6) was used in the last step. The probability $Q_\ell$ that the walks lasts exactly $\ell$ steps is accordingly given by

$$Q_\ell = P_\ell - P_{\ell+1} = \int_{-\infty}^{\infty} \mathrm{d}y\, P_\ell(y)\, F(y)^L\,\mathrm{d}y\,, \tag{4.8}$$

and thus the expected walk length reads

$$\mathbb{E}[\ell] = \sum_{\ell=0}^{\infty} \ell\, Q_\ell = \sum_{\ell=0}^{\infty} \ell\,(P_\ell - P_{\ell+1}) = \sum_{\ell=1}^{\infty} P_\ell\,. \tag{4.9}$$

Furthermore, the integrand of (4.8)

$$Q_\ell(x) = P_\ell(x)\, F(x)^L \tag{4.10}$$

is the joint probability (density) to have fitness $x$ in the $\ell$-th step and that the walk lasts exactly $\ell$ steps (conflated from the probabilities that the walk lasts at least $\ell$ steps and that the state with fitness $x$ is a local maximum). The marginal probability density of the final fitness $h$ is then given by

$$Q(x) = \sum_{\ell=0}^{\infty} Q_\ell(x)\,. \tag{4.11}$$

If the overall fitness distribution is uniform on $[0, 1]$, it will turn out that the limit

$$\lim_{L\to\infty} \frac{Q(1 - x'/L)}{L} = R(x') \tag{4.12}$$

leads to a non-degenerated function $R$ which is independent of $L$. This becomes useful for calculating the mean value of the final fitness. With the substitution $x = 1 - x'/L$ one finds

$$\begin{aligned} \mathbb{E}[h] &= \int_0^1 x\, Q(x)\,\mathrm{d}x = \frac{1}{L}\int_0^L \left(1 - \frac{x'}{L}\right) Q\left(1 - \frac{x'}{L}\right)\mathrm{d}x' \\ &= 1 - \frac{1}{L^2}\int_0^L x' Q\left(1 - \frac{x'}{L}\right)\mathrm{d}x' \end{aligned} \tag{4.13}$$

and hence

$$L\,(1-\mathbb{E}[h]) = \int_0^L x'\,\frac{Q(1-x'/L)}{L}\,\mathrm{d}x' \xrightarrow{L\to\infty} \int_0^\infty x'\,R(x')\,\mathrm{d}x'\,. \tag{4.14}$$

The leading order behavior of $\mathbb{E}[h]$ is therefore given by $\mathbb{E}[h] = 1-\alpha/L$, where $\alpha$ is given by the right-hand side of equation (4.14).

Unless otherwise stated, it is assumed in this section that fitness values are distributed uniformly on $[0,1]$. This happens without loss of generality for greedy, random and reluctant AWs, since only the order of random variables matters and not their actual value. Therefore, the walk length distribution $Q_\ell$ is independent of the overall fitness distribution while the height distribution $Q(x)$ can be obtained for the general case by a rather simple transformation of the uniform case. The latter will be discussed later in some more detail. For natural adaptive walks, however, the underlying fitness distribution affects the actual dynamics. In the following, the distribution and mean value of walk length $\ell$ and height $h$ for greedy, random and reluctant adaptive walks will be derived. A summary of the results can be found in table 2 on page 70 (see also figures 4.1 and 4.2). For the sake of completeness, known results for the natural adaptive walk will also be presented briefly.

### 4.1.1. Greedy Walks

For greedy walks, the transition matrix $\gamma(y \to x)$ is simply the probability density of the largest of $L$ random variables, but restricted to the region where $x > y$. This leads to

$$\gamma(y \to x) = L\,x^{L-1}\,. \tag{4.15}$$

The recurrence relation becomes

$$P_{\ell+1}(x) = L\int_0^x P_\ell(y)\,x^{L-1}\,\mathrm{d}y\,, \tag{4.16}$$

which has the solution

$$P_\ell(x) = \frac{\left(L\,x^L\right)^\ell}{\prod_{k=1}^{\ell-1}(kL+1)}\,. \tag{4.17}$$

The walk length distribution is accordingly given by

$$Q_\ell = \int_0^1 P_\ell(x)\,x^L\,\mathrm{d}x = \frac{L^\ell}{[(\ell+1)L+1]\prod_{k=1}^{\ell-1}(kL+1)}\,. \tag{4.18}$$

Using that $L/(kL+1) \approx 1/k$ for large $L$ yields

$$Q_\ell \approx \frac{\ell}{(\ell+1)!}\,, \tag{4.19}$$

i.e., the $L$-dependence vanishes for large $L$. Note that this result can also be obtained without solving the recurrence relation, as shown in [70]: If the walker takes a path $\sigma_0 \to \sigma_1 \to \ldots \to \sigma_\ell$, the genotype $\sigma_{i+1}$ has by definition of the greedy walk the largest fitness of the neighborhood of $\sigma_i$. Neglecting that the initial genotype might be a local maximum, a walker takes at least $\ell$ steps if the largest fitness of the first $\ell$ neighborhoods is in ascending order, i.e., $P_\ell \approx 1/\ell!$ due to their independence. Equation (4.19) is then obtained via $Q_\ell = P_\ell - P_{\ell-1}$. The fact that $Q_\ell$ is independent of $L$ implies also that the mean walk length

$$\mathbb{E}[\ell] = \sum_{\ell=0}^{\infty} \ell\, Q_\ell = \sum_{\ell=1}^{\infty} P_\ell \approx \sum_{\ell=1}^{\infty} \frac{1}{\ell!} = e - 1 = 1.7182\ldots \tag{4.20}$$

converges to a constant.

It follows from equation (4.11) that the walk height distribution is given by

$$Q(x) = \sum_{\ell=0}^{\infty} \frac{L^\ell\, x^{(\ell+1)L}}{\prod_{k=1}^{\ell-1}(kL+1)}\,. \tag{4.21}$$

In order to evaluate the sum, one could use the approximation $L/(kL+1) \approx 1/k$ again that leads to $Q(x) \approx x^L + Lx^{2L}\exp(x^L)$. Unfortunately, this expression is not normalized. However, a quite similar but normalized expression arises if one uses the fact that an adaptive walker terminates at the largest of all random variables it encounters during the walk. If the walk lasts $\ell$ steps, the final fitness is the largest of about $L(\ell+1)$ random variables which has probability density $L(\ell+1)\,x^{L(\ell+1)-1}$. Averaging over $\ell$ according to equation (4.19) yields

$$Q(x) \approx \sum_{\ell=0}^{\infty} \frac{L\,\ell\,(\ell+1)\,x^{L(\ell+1)-1}}{(\ell+1)!} = L\,x^{2L-1}\,e^{x^L}\,. \tag{4.22}$$

The limiting function $R(x')$ reads

$$R(x') = \lim_{L\to\infty} \frac{Q(1-x'/L)}{L} = e^{-2x'}\,e^{e^{-x'}}\,. \tag{4.23}$$

It follows with equation (4.14) that

$$\mathbb{E}[h] \approx 1 - \frac{\alpha_{\text{greed}}}{L}\,, \tag{4.24}$$

where

$$\alpha_{\text{greed}} = \int_0^\infty x'\,e^{-2x'}\,e^{e^{-x'}}\,\mathrm{d}x' = 0.4003\ldots\,. \tag{4.25}$$

A slightly different derivation for this constant can be found in [37].

### 4.1.2. Random Adaptive Walks

The transition matrix $\gamma(y \to x)$ for random AWs is the probability density of a uniform random variable on $[y, 1]$ times the probability that a genotype with fitness $y$ is not a local maximum. This leads to

$$\gamma(y \to x) = \frac{1 - y^L}{1 - y} \tag{4.26}$$

and

$$P_{\ell+1}(x) = \int_0^x \frac{1 - y^L}{1 - y} P_\ell(y) \, \mathrm{d}y \,. \tag{4.27}$$

Following Flyvbjerg and Lautrup [66], one can define

$$H(y) = \sum_{k=1}^{L} \frac{y^k}{k} \,, \tag{4.28}$$

which has the nice property that

$$\mathrm{d}H = \left( \sum_{k=1}^{L} y^{k-1} \right) \mathrm{d}y = \frac{1 - y^L}{1 - y} \, \mathrm{d}y \,. \tag{4.29}$$

It will also be needed that

$$H\left(1 - \frac{x}{L}\right) = \log L + \gamma_{\mathrm{E}} - \mathrm{Ein}(x) + \mathcal{O}\left(\frac{1}{L}\right) , \tag{4.30}$$

where $\gamma_{\mathrm{E}} = 0.5772\ldots$ is the Euler-Mascheroni constant and

$$\mathrm{Ein}(x) = \int_0^x \frac{1 - e^{-t}}{t} \, \mathrm{d}x \,, \tag{4.31}$$

which in turn has the property that

$$\int_0^\infty e^{-x - \mathrm{Ein}(x)} \, \mathrm{d}x = e^{-\gamma_{\mathrm{E}}} \,. \tag{4.32}$$

Equation (4.27) written in terms of $H$ reads

$$P_{\ell+1}(H) = \int_0^H P_\ell(H') \, \mathrm{d}H' \tag{4.33}$$

and has the rather trivial solution $P_\ell(H) = H^\ell / \ell!$ or, in terms of $x$,

$$P_\ell(x) = \frac{1}{\ell!} \left( \sum_{k=1}^{L} \frac{x^k}{k} \right)^\ell = \frac{1}{\ell!} H(x)^\ell \,. \tag{4.34}$$

The walk length distribution reads then

$$Q_\ell = \frac{1}{\ell!}\int_0^1 x^L\, H(x)^\ell \,\mathrm{d}x\,. \tag{4.35}$$

This expression is a bit hard to evaluate directly, but its generating function

$$\hat{Q}(\lambda) = \sum_{\ell=0}^{\infty} \lambda^\ell\, Q_\ell = \int_0^1 x^L\, e^{\lambda\, H(x)}\,\mathrm{d}x \tag{4.36}$$

is very useful. With the substitution $y = 1 - x/L$ and equation (4.30), one can approximate the function by

$$\begin{aligned}\hat{Q}(\lambda) &= \frac{1}{L}\int_0^L \Big(1-\frac{y}{L}\Big)^L e^{\lambda\,\mathrm{Ein}(1-y/L)}\,\mathrm{d}y \\ &\approx \frac{1}{L}\int_0^\infty e^{-y+\lambda(\log L+\gamma_\mathrm{E}-\mathrm{Ein}(y))}\,\mathrm{d}y \\ &= L^{\lambda-1}\int_0^\infty e^{-y+\lambda[\gamma_\mathrm{E}-\mathrm{Ein}(y)]}\,\mathrm{d}y\,.\end{aligned} \tag{4.37}$$

The leading order of the length distribution turns out to be a Poisson distribution with parameter $\log L$:

$$\begin{aligned}Q_\ell &= \frac{1}{\ell!}\frac{\mathrm{d}^\ell \hat{Q}(\lambda)}{\mathrm{d}\lambda^\ell}\bigg|_{\lambda=0} \approx \frac{1}{L\,\ell!}\int_0^\infty [\log L+\gamma_\mathrm{E}-\mathrm{Ein}(y)]^\ell e^{-y}\,\mathrm{d}y \\ &= \frac{1}{L\,\ell!}\Big\{(\log L)^\ell + \mathcal{O}\Big[(\log L)^{\ell-1}\Big]\Big\}\,.\end{aligned} \tag{4.38}$$

The next order correction to the first moment of the actual distribution, which is just an additive shift, can be computed as well. Equation (4.32) yields

$$\begin{aligned}\mathbb{E}[\ell] = \sum_{\ell=0}^{\infty}\ell\, Q_\ell = \frac{\mathrm{d}\hat{Q}(\lambda)}{\mathrm{d}\lambda}\bigg|_{\lambda=1} &\approx \int_0^\infty [\log L+\gamma_\mathrm{E}-\mathrm{Ein}(y)]e^{-y+\gamma_\mathrm{E}-\mathrm{Ein}(y)}\,\mathrm{d}y \\ = \log L + \int_0^\infty [\gamma_\mathrm{E}-\mathrm{Ein}(y)]e^{-y+\gamma_\mathrm{E}-\mathrm{Ein}(y)}\,\mathrm{d}y &= \log L + c\,,\end{aligned} \tag{4.39}$$

where $c \approx 0.0991$. A similar result was obtained in [65] which states that

$$\mathbb{E}[\ell] \approx \log(L) + c + \log(1-x_0) + 1\,, \tag{4.40}$$

where $x_0$ is the fitness of the initial genotype. Equation (4.39) can be obtained from (4.40) by averaging over $x_0$. Note also that for $x_0 = 0$, the first step will lead to a random fitness that is uniformly distributed on $[0,1]$ and hence the average walk will be one step longer than a walk that started with uniformly distributed fitness in the first place.

Now consider the height distribution. According to equation (4.11) it is given by

$$Q(x) = x^L \sum_{\ell=0}^{\infty} P_\ell(x) = x^L e^{H(x)} \,. \tag{4.41}$$

Using equation (4.30), the limiting function $R(x')$ can be computed and reads

$$R(x') = \lim_{L\to\infty} \frac{Q(1-x'/L)}{L} = e^{-x'+\gamma_{\mathrm{E}}-\mathrm{Ein}(y)} \,. \tag{4.42}$$

With this, the mean height can be approximated by

$$\mathbb{E}[h] \approx 1 - \frac{\alpha_{\mathrm{rnd}}}{L} \,, \tag{4.43}$$

where

$$\alpha_{\mathrm{rnd}} = \int_0^\infty y\, e^{-y+\gamma_{\mathrm{E}}-\mathrm{Ein}(y)}\, \mathrm{d}y = 0.6243\ldots \,. \tag{4.44}$$

Note that this result was also obtained in [65] with a similar ansatz.

### 4.1.3. Reluctant Walks

For reluctant walks, $\gamma(y \to x)$ is the probability density of the smallest of $L$ random variables that are larger than $y$. The number $k$ of random variables that are larger than $y$ is binomial distributed. Given that $k$ random variables are larger than $y$, the density $\gamma_k(y \to x)$ of the smallest of them is given by

$$\gamma_k(y \to x) = \frac{k}{1-y}\left(1 - \frac{x-y}{1-y}\right)^{k-1} \,, \tag{4.45}$$

which yields

$$\begin{aligned}
\gamma(y \to x) &= \sum_{k=1}^{L} \binom{L}{k} (1-y)^k\, y^{L-k} \gamma_k(y \to x) \\
&= \sum_{k=1}^{L} \binom{L}{k} k\, (1-x)^{k-1}\, y^{L-k} \\
&= L\,(1-x+y)^{L-1} \,.
\end{aligned} \tag{4.46}$$

The recurrence relation (4.7) becomes

$$P_{\ell+1}(x) = L \int_0^x P_\ell(y)(1-x+y)^{L-1}\, \mathrm{d}y \,. \tag{4.47}$$

Reluctant walks, in contrast to the other walk types, are very sensitive to the initial fitness. It will turn out useful to condition the starting fitness to be $x_0 = 1 - x_0'/L$,

i.e., $P_0(x) = \delta(x - x_0)$ and $P_\ell(x) = 0$ for $x < x_0$. Furthermore, let $x = 1 - x'/L$ and $y = 1 - y'/L$ with $0 \le x', y', x_0' \le L$ and $\tilde{P}_\ell(x') = P_\ell(1 - x'/L)$. Then, equation (4.47) can be written in terms of $\tilde{P}_\ell(x')$ as

$$\tilde{P}_{\ell+1}(x') = \int_{x'}^{x_0'} \tilde{P}_\ell(y') \left(1 - \frac{y' - x'}{L}\right)^{L-1} \mathrm{d}y' \,. \tag{4.48}$$

For large $L$, one can approximate the second factor of the integrand by an exponential function, which reads

$$\tilde{P}_{\ell+1}(x') \approx \int_{x'}^{x_0'} \tilde{P}_\ell(y')\, e^{x'-y'}\, \mathrm{d}y' \,. \tag{4.49}$$

It is straightforward to check that

$$\tilde{P}_\ell(x') = \frac{(x_0' - x')^{\ell-1}}{(\ell-1)!}\, e^{x'-x_0'}\, I_{[0,x_0']}(x') \tag{4.50}$$

solves the recurrence relation (4.49).

Applying the same transformation and approximation to equation (4.10) yields

$$\tilde{Q}_\ell(x') = \tilde{P}_\ell(x')\, e^{-x'} = \begin{cases} \frac{e^{-x_0'}(x_0'-x')^{\ell-1}}{(\ell-1)!}\, I_{[0,x_0']}(x') \,, & \text{if } \ell > 0, \\ \delta(x' - x_0')\, e^{-x'} \,, & \text{if } \ell = 0. \end{cases} \tag{4.51}$$

Unfortunately, the case study in (4.51) is needed to ensure the normalization of the height distribution. The walk length distribution is then given by

$$Q_\ell = \int_0^{x_0'} \tilde{Q}_\ell(y')\, \mathrm{d}y' = \int_0^{x_0'} \frac{e^{-x_0'}(x_0' - x')^{\ell-1}}{(\ell-1)!}\, \mathrm{d}y' = \frac{e^{-x_0'}\,(x_0')^\ell}{\ell!} \,, \tag{4.52}$$

i.e., the number of steps is Poisson distributed with parameter $x_0' = L(1 - x_0)$. If the initial fitness is chosen randomly from $[0, 1]$, the mean walk length is obviously given by $\mathbb{E}[\ell] = L/2$. The corresponding length distribution can be obtained by integrating over $x_0'$ which leads to

$$Q_{\ell,\mathrm{rnd}} = \frac{1}{L} \int_0^L \tilde{Q}_\ell \, \mathrm{d}x_0' = \frac{1}{L} \left(1 - \frac{\Gamma(\ell+1, L)}{\ell!}\right), \tag{4.53}$$

where $\Gamma(a, b) = \int_b^\infty t^{a-1} e^{-t}\, \mathrm{d}t$ is the incomplete Gamma function.

The density of the (transformed) height is given by

$$\begin{aligned} \tilde{Q}(x') &= \sum_{\ell=0}^{\infty} \tilde{Q}_\ell(x') = Q_0(x') + \sum_{\ell=1}^{\infty} \tilde{Q}_\ell(x') \\ &= \delta(x' - x_0') e^{-x'} + I_{[0,x_0']}(x') \sum_{\ell=1}^{\infty} \frac{e^{-x_0'}\,(x_0')^\ell}{\ell!} \\ &= e^{-x'} \Big[ I_{[0,x_0']}(x') + \delta(x' - x_0') \Big] \end{aligned} \tag{4.54}$$

and the transformation from $x'$ back to $x$ yields

$$Q(x) = e^{-L(1-x)} \left[ L \cdot I_{[x_0,1]}(x) + \delta(x - x_0) \right] . \tag{4.55}$$

Finally, the limiting function $R(x')$ is simply given by

$$R(x') = \lim_{L\to\infty} \frac{Q(1 - x'/L)}{L} = e^{-x'} \tag{4.56}$$

which leads with equation (4.14) to

$$\mathbb{E}[h] \approx 1 - \frac{1}{L} \int_0^\infty x' \, e^{-x'} \, \mathrm{d}x' = 1 - \frac{1}{L} \, , \tag{4.57}$$

i.e., $\alpha_{\text{reluc}} = 1$. Interestingly, this corresponds to the fitness of a randomly chosen local maximum.

### 4.1.4. Approximations for Non-Uniform Fitness Distributions

As already mentioned, the choice of the fitness distribution has no influence on the dynamics (with natural AWs being an exception). The length distribution $Q_\ell$ is therefore completely independent of the choice, but obviously the height distribution changes. Let $Q(x)$ denote the probability density of the height in the uniform case and $Q^*(x)$ the density for the general case where the overall fitness is distributed according to a cumulative distribution function $F(x)$ with corresponding density $f(x)$. Then the transformation from $Q$ to $Q^*$ reads

$$Q^*(x) = f(x) \, Q[F(x)] \, . \tag{4.58}$$

In some cases one might be only interested in the mean value $\mathbb{E}[h]$ which is given by

$$\mathbb{E}[h] = \int_{-\infty}^{\infty} x \, Q^*(x) \, \mathrm{d}x = \int_0^1 F^{-1}(x) \, Q(x) \, \mathrm{d}x \, . \tag{4.59}$$

Since it is a bit cumbersome to evaluate this integral, it would be nice to have the scaling behavior in a form as simple as equation (4.14). This depends especially on the tail of the underlying fitness distribution $F$ since the height distribution $Q(x)$ has most of its weight for $x$ close to one. It is known from extreme value theory [75] that the tail behavior of a distribution can be nicely represented by the generalized Pareto distribution (GPD) with CDF

$$F_\kappa(x) = \begin{cases} 1 - (1 + \kappa x)^{-1/\kappa} & \text{for } \kappa \neq 0, \\ 1 - e^{-x} & \text{for } \kappa = 0. \end{cases} \tag{4.60}$$

The support is $\mathbb{R}^+$ for $\kappa \geq 0$ and $[0, -1/\kappa]$ for $\kappa < 0$. Depending on the value of $\kappa$, the distributions can be member of either of the three universality classes of extreme value theory: For $\kappa > 0$, the tail decays as a power law corresponding to the Fréchet class,

| Quantity | Greedy AW | Random AW | Reluctant AW |
|---|---|---|---|
| $Q_\ell$ | $\dfrac{\ell}{(\ell+1)!}$ | $\dfrac{(\log L)^\ell}{L\,\ell!}$ | $\dfrac{1}{L}\left(1-\dfrac{\Gamma(\ell+1,L)}{\ell!}\right)$ |
| $Q(x)$ | $L\,e^{x^L}x^{2L-1}$ | $x^L e^{H(x)}$ | $L\,e^{-L(1-x)}$ |
| $\mathbb{E}[\ell]$ | $e-1$ | $\log L+0.1$ | $L/2$ |
| $L(1-\mathbb{E}[h])$ | $\alpha_{\text{greed}}=0.4002\ldots$ | $\alpha_{\text{rnd}}=0.6243\ldots$ | $\alpha_{\text{reluc}}=1$ |

**Table 2.** Approximations of adaptive walk properties on the HoC model with $\mathcal{U}(0,1)$ distributed fitness and random initial condition. The derivation of all quantities can be found in the text.

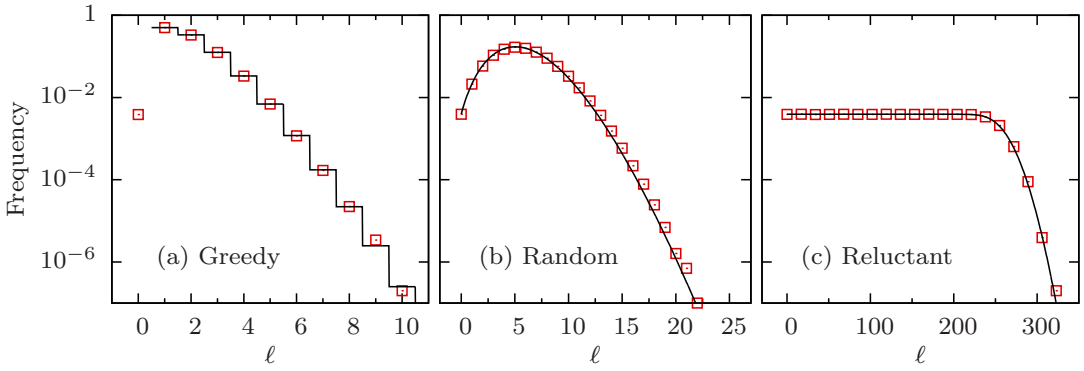



**Figure 4.1.** Walk length distribution for (a) greedy, (b) random and (c) reluctant adaptive walks on an $L=256$ HoC landscape with randomly chosen initial fitness. Symbols correspond to numerical results, solid lines to analytical results listed in table 2.

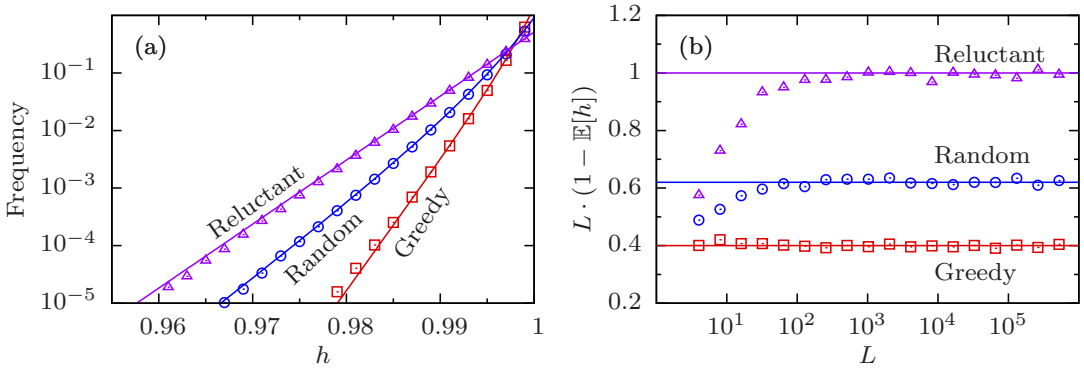



**Figure 4.2.** (a) Walk height distribution for greedy, random and reluctant adaptive walks on an $L=256$ HoC landscape with randomly chosen initial fitness. Symbols correspond to numerical results, solid lines to analytical results listed in table 2. (b) Rescaled average walk height for the same walk types. Symbols correspond to numerical results, solid lines to the constants $\alpha_{\text{reluc}}$, $\alpha_{\text{rnd}}$ and $\alpha_{\text{greed}}$.

for $\kappa = 0$ it decays exponentially corresponding to the Gumbel class and for $\kappa < 0$ it is bounded which corresponds to the Weibull class. However, only the case where $\kappa < 1$ is considered, because otherwise the GDP and accordingly the height distribution have an infinite mean value. In order to evaluate the right-hand side of (4.59), one needs the quantile function which is given by

$$F_\kappa^{-1}(x) = \begin{cases} \frac{(1-x)^{-\kappa}-1}{\kappa} & \text{for } \kappa \neq 0, \\ -\log(1-x) & \text{for } \kappa = 0. \end{cases} \tag{4.61}$$

One finds for $\kappa \neq 0$ and $\kappa < 1$ that

$$\begin{aligned} \mathbb{E}[h] &= \int_0^1 F_\kappa^{-1}(x)\, Q(x)\, \mathrm{d}x = \frac{1}{\kappa\, L} \int_0^L \left[ \left(\frac{y}{L}\right)^{-\kappa} - 1 \right] Q\Big(1 - \frac{y}{L}\Big)\, \mathrm{d}y \\ &\approx \frac{1}{\kappa} \int_0^\infty \left[ \left(\frac{y}{L}\right)^{-\kappa} - 1 \right] R(y)\, \mathrm{d}y = \frac{1}{\kappa}(L^\kappa \beta(\kappa) - 1)\,, \end{aligned} \tag{4.62}$$

where

$$\beta(\kappa) = \int_0^\infty y^{-\kappa}\, R(y)\, \mathrm{d}y\,.$$

Analogously, for $\kappa = 0$ one finds

$$\mathbb{E}[h] = \frac{1}{L} \int_0^L \log\left(\frac{L}{y}\right) Q\Big(1 - \frac{y}{L}\Big)\, \mathrm{d}y \approx \int_0^\infty \log\left(\frac{L}{y}\right) R(y)\, \mathrm{d}y = \log L - \beta(0) \tag{4.63}$$

where

$$\beta(0) = \int_0^\infty \log(y)\, R(y)\, \mathrm{d}y\,.$$

Unfortunately, the coefficient $\beta(\kappa)$ can only be evaluated in simple closed form in case of reluctant AWs for which $R(x) = \exp(-x)$. This yields

$$\beta(\kappa) = \begin{cases} \Gamma(1-\kappa) & \text{for } \kappa \neq 0, \\ \gamma_\mathrm{E} & \text{for } \kappa = 0\,. \end{cases}$$

Now a rather loose but more general approximation for the mean of $h$ will be considered. A similar derivation can be found in [37]. Note that one can write the right-hand side of (4.63) for $\kappa = 0$ and reluctant AWs as

$$\log L - \gamma_\mathrm{E} = F_0^{-1}\left(1 - \frac{\alpha_\mathrm{reluc} \cdot \exp(-\gamma_\mathrm{E})}{L}\right).$$

This might seem a bit odd, especially since $\alpha_\mathrm{reluc} = 1$, but it turns out numerically that

$$\mathbb{E}[h] \approx F^{-1}\left(1 - \frac{\alpha\, \exp(-\gamma_\mathrm{E})}{L}\right), \tag{4.64}$$

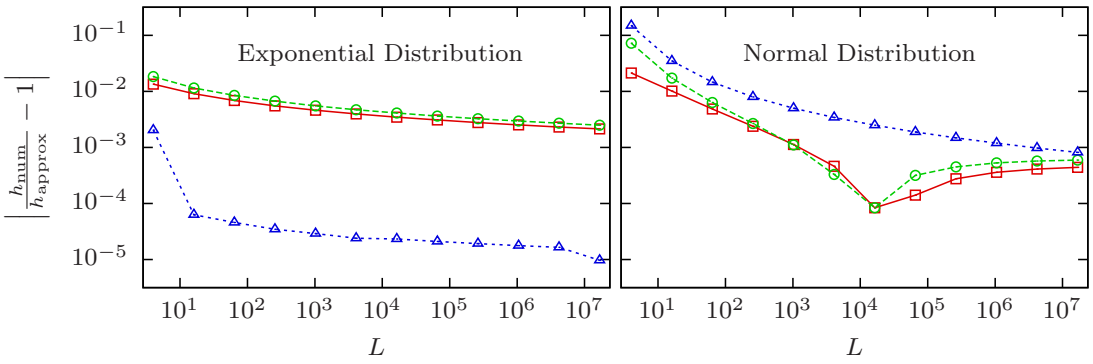



**Figure 4.3.** Relative error of the approximation $h_{\text{approx}}$ of the mean height given by equation (4.64). The "correct" value $h_{\text{num}}$ is obtained from the numerical integration of equation (4.59). Red squares correspond to greedy walks, green circles to random walks and blue triangles to reluctant walks. The minima in the right panel are due to a sign change of $h_{\text{approx}} - h_{\text{num}}$.

where $\alpha = \lim_{L\to\infty} L(1 - \mathbb{E}[h])$, is also a good approximation for walk types other than reluctant and even for other distributions in the Gumbel class with distribution function $F$. In particular, this applies to Gaussian distributions that will be used in section 4.3. It is not necessarily expected that (4.64) gives the correct asymptotic behavior (except for reluctant walks on an exponentially distributed landscape), but as figure 4.3 shows, the relative error is smaller than 1% for sufficiently large values of $L$.

### 4.1.5. Natural Adaptive Walks

Up to now, the natural adaptive walk does not fit quite well in the previous scheme as neither $Q_\ell$ nor $Q(x)$ are known exactly for general distributions. This is because the dynamics is significantly influenced by the choice of the fitness distribution and the recursion equation (4.7) becomes difficult to solve if the transition probabilities for that walk type are inserted. Nevertheless, some results for the generalized Pareto distribution were obtained [76, 77].

In order to formulate the recursion, one can replace the sum in the denominator of equation (4.2) by an integral for $\kappa < 1$ [78] such that

$$\gamma(y \to x) = \frac{(x-y)\,f(x)}{\int_y^\infty (z-y)\,f(z)\,\mathrm{d}z}\left[1 - F(y)^L\right], \tag{4.65}$$

where $f$ and $F$ are probability density and cumulative distribution function of the fitness. The second factor takes explicitly into account that a step is only possible if the genotype with fitness $y$ is not a local optimum. Equation (4.7) was studied in [77] with this transition matrix, where the first three central moments of the length distribution $Q_\ell$

have been carried out to leading order in $L$. They read

$$\begin{aligned}
\mathbb{E}[\ell] &= \frac{1-\kappa}{2-\kappa}\log L\,,\\
\mathbb{E}\Big[\big(\ell-\mathbb{E}[\ell]\big)^2\Big] &= \frac{(1-\kappa)\,(2-2\kappa+\kappa^2)}{(2-\kappa)^3}\log L\,,\\
\mathbb{E}\Big[\big(\ell-\mathbb{E}[\ell]\big)^3\Big] &= \frac{(1-\kappa)\,(4-8\kappa+6\kappa^2-2\kappa^3+\kappa^4)}{(2-\kappa)^5}\log L\,.
\end{aligned}$$

Like for random adaptive walks, the expected walk length increases logarithmically with $L$, but with a $\kappa$-dependent prefactor that decreases with increasing $\kappa$. For $\kappa \to -\infty$, the prefactor converges to 1 while it vanishes for $\kappa \to 1$ and hence the walk length behavior interpolates between the random and greedy case with respect to the fitness distribution's tail behavior. Intuitively, this happens because the transition probabilities (4.2) are dominated by the largest fitness in case of a heavy tailed distribution and hence resemble greedy behavior. If, on the other hand, the distribution is very narrow, all transition probabilities would usually have similar values such that a random neighbor is chosen in each step.

## 4.2. Adaptive Walks on the RMF Landscape

Like HoC landscapes, the RMF landscape can be approximated well on a tree in order to get a recurrence relation similar to that in (4.7), but it is hard to solve for arbitrary distributions. In general, the adaptive walk behavior is governed by the underlying distribution of the random part of the landscape, even in case of walk types that are not affected by the fitness distribution on the HoC landscape.

### 4.2.1. Greedy Adaptive Walks

The case of greedy adaptive walks with Gumbel distributed random part was studied in [79] (see also [80] for further studies). It was assumed that the reference sequence $\tilde{\sigma}$ is the initial genotype of the walk, i.e., the potentially least fit genotype. Similar to the approach by Orr [70] for greedy walks on the HoC landscape, one can estimate $P_\ell$ by the fact that a walk lasts at least $\ell$ steps if the largest fitness among the first $\ell$ neighborhoods that the walker "sees" are ordered. Due to the properties of the Gumbel distribution, the probability for this can be obtained from the ordering probability given by equation (3.76). Interestingly, the result can also be expressed similarly to the one for the HoC model by using the so-called $q$-analogue [81]. A $q$-number $[n]_q$ is defined as

$$[n]_q = \frac{1-q^n}{1-q}\,. \tag{4.66}$$

Many other functions can then be redefined using these numbers instead of integers in their definition. For instance, the $q$-factorial reads

$$[n]_q! = \prod_{k=1}^{n} [k]_q \,. \tag{4.67}$$

It was argued before that $P_\ell = 1/\ell!$ for greedy walks on the HoC landscape. Analogously, the probability $P_\ell$ that a greedy walks lasts at least $\ell$ on the RMF landscape is given by the reciprocal $q$-factorial

$$P_\ell = \frac{1}{[\ell]_{\exp(-s)}!} = \prod_{k=1}^{\ell} \frac{1}{[k]_{\exp(-s)}} \,. \tag{4.68}$$

With this, the mean walk length is given by

$$\mathbb{E}[\ell] = \exp_{\exp(-s)}(1) - 1 \,, \tag{4.69}$$

where

$$\exp_q(x) = \sum_{k=0}^{\infty} \frac{x^k}{[k]_q!}$$

is the $q$-analogue of the exponential function in nice analogy to equation (4.20). This also reveals that the walk length is still bounded for $L \to \infty$, which means that even though accessible paths to the global optimum exist with probability 1 according to equation (3.89), a greedy adaptive walker will not take these paths.

### 4.2.2. Random Adaptive Walks

In case of random adaptive walks, the counterpart to equation (4.27) reads

$$P_{\ell+1}(x) = \int_{-\infty}^{x+s} \frac{1 - F(y-s)^{L-\ell}}{1 - F(y-s)} \, f(x) \, P_\ell(y) \, \mathrm{d}y \,, \tag{4.70}$$

which was studied for a standard exponential distribution in [82]. Again, the walker starts from the reference sequence $\tilde{\sigma}$. The main result is that there is a phase transition of the mean walk length with respect to $s$. More precisely, one has

$$\mathbb{E}[\ell] \propto \begin{cases} \log(L)/(1-s) \,, & \text{for } s < 1, \\ \log(L)^2 \,, & \text{for } s = 1, \\ \mathcal{O}(L) \,, & \text{for } s > 1. \end{cases} \tag{4.71}$$

This confirms the intuition that the behavior on an RMF landscape changes when the fluctuations are of the same order as the slope $s$. Moreover, if the distributions tail decays more slowly than exponentially, the walk length is linear in $L$, while for tails decaying faster than exponentially, the walk length grows logarithmically with $L$. Note that this coincides nicely with the change of the behavior of local maxima according to equation (2.34).

## 4.3. Adaptive Walks on the NK Landscape

### 4.3.1. Block Interactions

As shown before in section 2.4.3 and 3.5.1, a blockwise interaction pattern facilitates analytical studies since many quantities can be derived from known results of the HoC model. Adaptive walks proceed independently within each block as pointed out in [39] for random adaptive walks. In fact, this is also the case for a rather general family of adaptive walks, namely if the transition probabilities are only a function of fitness differences.

In that case, the transition probabilities can be written as

$$p(\sigma \to \tau) = \frac{\xi[w(\tau) - w(\sigma)]}{\sum_{\sigma' \in \mathcal{U}_\sigma^+} \xi[w(\sigma') - w(\sigma)]}\,, \tag{4.72}$$

where $\xi\colon \mathbb{R}^+ \to \mathbb{R}^+$ is an arbitrary function and hence, according to equation (4.5), all previously defined walk types can be obtained for suitable choices of $\xi$. In order to show that the whole AW can be treated as a set of walks through HoC landscapes, some special notation related to the switching between the whole landscape and the sub-landscapes defined by the different blocks is needed. For $\sigma \in \mathbb{H}_2^L$, let $\sigma^{(b)}$ be the subsequence of length $K$ corresponding to the $b$-th block and $w_b\big(\sigma^{(b)}\big)$ the fitness landscape of that block, i.e.,

$$w(\sigma) = \sum_{b=1}^{L/K} w_b\Big(\sigma^{(b)}\Big)\,.$$

Furthermore, let $b(\sigma, \tau)$ be the block in which a mutation has to occur to step from $\sigma$ to $\tau$ (or vice versa), $B(\sigma)$ the random variable that takes the value of the block in which an adaptive step from $\sigma$ takes place and $\mathcal{U}_b^+(\sigma)$ the set of neighbors of $\sigma$ that can be reached by a beneficial mutation in the $b$-th block. Then the probability $p(\sigma \to \tau \mid b)$ for an adaptive step, conditioned on taking place in the $b$-th block with $b = b(\sigma, \tau)$, is given by

$$\begin{aligned}
p(\sigma \to \tau \mid b) &= \frac{p(\sigma \to \tau)}{\mathbb{P}[B(\sigma) = b]} = p(\sigma \to \tau) \frac{\sum_{\sigma' \in \mathcal{U}_\sigma^+} p(\sigma \to \sigma')}{\sum_{\sigma' \in \mathcal{U}_b^+(\sigma)} p(\sigma \to \sigma')} \\
&= \frac{\xi[w(\tau) - w(\sigma)]}{\sum_{\sigma' \in \mathcal{U}_\sigma^+} \xi[w(\sigma') - w(\sigma)]} \, \frac{\sum_{\sigma' \in \mathcal{U}_\sigma^+} \xi[w(\sigma') - w(\sigma)]}{\sum_{\sigma' \in \mathcal{U}_b^+(\sigma)} \xi[w(\sigma') - w(\sigma)]} \\
&= \frac{\xi[w(\tau) - w(\sigma)]}{\sum_{\sigma' \in \mathcal{U}_b^+(\sigma)} \xi[w(\sigma') - w(\sigma)]} \\
&= \frac{\xi\big[w_b\big(\tau^{(b)}\big) - w_b\big(\sigma^{(b)}\big)\big]}{\sum_{\sigma' \in \mathcal{U}_b^+(\sigma)} \xi\big[w_b\big(\sigma'^{(b)}\big) - w_b\big(\sigma^{(b)}\big)\big]} = \tilde{p}\big(\sigma^{(b)} \to \tau^{(b)}\big)\,.
\end{aligned} \tag{4.73}$$

Note that the last line has the same form as (4.72), but depends solely on the corresponding block. In other words, it is the probability $\tilde{p}(\tilde{\sigma} \to \tilde{\tau})$ that an adaptive

step from $\tilde{\sigma} = \sigma^{(b)}$ to $\tilde{\tau} = \tau^{(b)}$ would have in this HoC sub-landscape of size $K$. In the following it is convenient to define $\tilde{p}(\tilde{\sigma} \to \tilde{\sigma}) = 1$ such that the probability for a whole path can be written as

$$\begin{aligned}\mathbb{P}[\sigma_0 \to \sigma_1 \to \ldots \to \sigma_\ell] &= \prod_{i=1}^{\ell} p(\sigma_{i-1} \to \sigma_i) \\ &= \prod_{b=1}^{L/K} \underbrace{\left[\prod_{i=1}^{\ell} \tilde{p}\big(\sigma_{i-1}^{(b)} \to \sigma_i^{(b)}\big)\right]}_{P_{\text{paths}}} \cdot \underbrace{\left[\prod_{i=1}^{\ell} \mathbb{P}[B(\sigma_{i-1}) = b(\sigma_{i-1}, \sigma_i)]\right]}_{P_{\text{order}}} . \end{aligned} \tag{4.74}$$

The first expression, $P_{\text{paths}}$, is nothing but the probability that a a certain path is taken through a HoC landscape of size $K$. $P_{\text{order}}$ is the probability that the blocks in which mutations occur have a certain order. However, this order has no influence on the walk length $\ell$ or height $h$. They are both additive in the sense that

$$\ell = \sum_{b=1}^{L/K} \ell_b \qquad \text{and} \qquad h = \sum_{b=1}^{L/K} h_b \,, \tag{4.75}$$

where $\ell_b$ and $h_b$ are length and height, respectively, of the adaptive walk through the $b$-th sub-landscape. As the AWs through the sub-landscapes behave as ordinary AWs through a $K$-dimensional HoC landscape, also $\ell_b$ and $h_b$ have the same distribution. Accordingly, the distribution of $\ell$ and $h$ is the $L/K$-fold convolution of length and height distribution, respectively, of a $K$-dimensional HoC landscape. As an example, the length distribution of the random adaptive walk is approximately given by a Poisson distribution with mean $L/K \cdot \log K$. It is noteworthy that this mean value already appeared in [42], but as an approximation for adjacent and random interaction patterns and not as an asymptotically exact expression for block interactions like in [39].

Rather than studying the full distribution, the focus will be on the mean value of length and height in the following. Obviously, equation (4.75) implies

$$\mathbb{E}[\ell] = \frac{L}{K} \mathbb{E}[\ell_{\text{HoC}}(K)] \,. \tag{4.76}$$

The values for $\mathbb{E}[\ell_{\text{HoC}}(K)]$ can be taken from table 2. An interesting observation at this point is that $\mathbb{E}[\ell] \approx L/2$ for reluctant walks, independent of $K$. Nevertheless, remember that the expression given in the table are only valid asymptotically for large values of $K$ due to the fact that they were obtained for a tree and that further approximations were made. Therefore, if the actual values of (4.76) are needed, small-$K$ corrections to $\mathbb{E}[\ell_{\text{HoC}}(K)]$ will be included that are proportional to $K^{-1}$ and $K^{-2}$ with coefficients obtained from a least square fit of simulation data.

With the same argument and equation (4.64) one finds

$$\mathbb{E}[h] = \frac{L}{K} \mathbb{E}[h_{\text{HoC}}(K)] \approx \frac{L}{K} F_{\text{b}}^{-1}\left(1 - \frac{\alpha \, \exp(-\gamma_{\text{E}})}{K}\right), \tag{4.77}$$

where $\alpha$ is the coefficient depending on the walk type and $F_\mathrm{b}$ is the distribution function of fitness within a block. As the fitness contributions are standard normal distributed throughout this section, the fitness of a block is normal distributed with variance $K$, i.e.,

$$F_\mathrm{b}^{-1}(x) = \sqrt{2}\,K\,\mathrm{erf}^{-1}(2x-1)\,, \tag{4.78}$$

where $\mathrm{erf}^{-1}$ is the inverse error function. The comparison of equations (4.76) and (4.77) with simulation data can be found in figure 4.5.

### 4.3.2. General Phenomenology

As shown in chapter 2 and 3, most quantities of the NK model can be tuned by the parameter $K$ and, in a more subtle way, by the choice of the interaction pattern. Therefore, it is a nice way to study the influence of landscape properties on the dynamics.

Simulations reveal that both mean length and height increase asymptotically linear with sequence length $L$ if $K$ is kept constant. This behavior is not very surprising: Increasing $L$ also increases the number of contributions to the fitness which in turn is proportional to the final fitness $h$. With regard to walk length $\ell$, linear behavior follows directly from equations (4.76) for block interaction, i.e., the number of independent blocks grows linearly with $L$. Other interactions schemes do not have this block structure in a strict sense, but when $L$ is sufficiently larger than $K$, epistatic interactions are short ranged and hence one can still find different areas of the genotype sequence that are largely independent, similar to the argument given in section 3.5. The number of these areas also grows linearly and hence the linear behavior of $\mathbb{E}[\ell]$. How good this argument is can be seen from the $y$-intercepts $b$ of linear regressions of the form $a\,L + b$ (see figure 4.4). If there is a strictly linear dependence between $\mathbb{E}[\ell]$ and $L$, $b$ would be zero. This is only the case for block interactions, but in most other cases, $b$ is of the same order of magnitude as the slope $a$, i.e., rather small. A notable exception is the combination of reluctant walks and random interaction. The slope is already very large ($a \approx 4.9$ in the example shown here), but the absolute value of the $y$-intercept ($b \approx -208$ in the example) is even many times larger, implying that a linear approximation fails if $L$ is of the same order as $K$.

Intuitively it is clear that an adaptive walk is longer when the fitness landscape is less rugged, which is mostly confirmed here: In agreement with other measures of ruggedness that were shown before, adaptive walk lengths increase with increasing rank of the interaction scheme. For most cases, the intuition is also true if the ruggedness is controlled by the landscape parameter $K$ rather than the interaction scheme, as shown in the left-hand side of figure 4.5. Reluctant walks are again an exception. Like mentioned before, the mean length is always $L/2$ for block interactions and is hence independent of $K$. For other interaction types, the $K$-dependence is even non-monotonic for fixed $L$. Especially the graph for random interactions shows a rather sharp maximum with a height that is several times larger than the system size. Reluctant walks show by far the highest susceptibility to different interaction patterns in terms of length.

With regard to the walk heights shown in figure 4.5(d)-(f), the same order as for lengths occurs: The largest fitness is reached with random interactions, the second

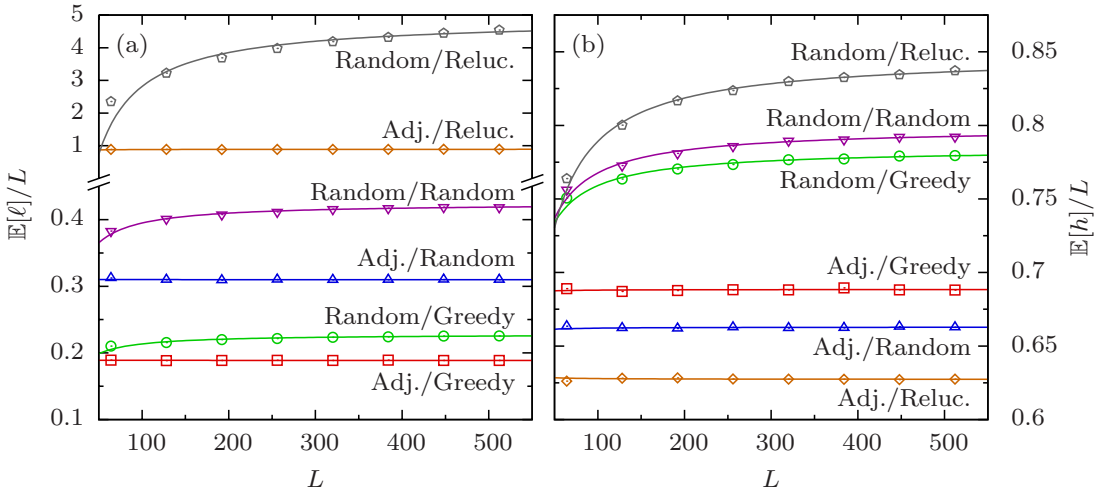



**Figure 4.4.** Averaged walk length (a) and height (b) on an NK landscape with constant $K = 8$. Symbols correspond to numerical results. Labels are of the form "Interaction scheme/Walk type". Lines correspond to linear regression of $\mathbb{E}[\ell]$ and $\mathbb{E}[h]$, respectively. The curvature of $\mathbb{E}[\ell]/L$ and $\mathbb{E}[h]/L$ is due to a non-zero $y$-intercept of the regression.

largest fitness with adjacent and the smallest with block interactions in accordance with figure 2.2(c). The dependence of $\mathbb{E}[h]$ on $K$, however, is non-monotonic with a maximum located at rather small values of $K$. This is in contrast to figure 2.2(d) where it is shown that the height of typical maxima decreases monotonically with $K$. The maximum might be due to a tradeoff between a decreased fitness of maxima and an increase of their number which might enhance the ability to find particularly fit ones. Like for the length, reluctant walks are more susceptible to the interaction scheme. Even though the difference might seem rather small, it has a big impact on the question which walk type reaches the highest fitness. This will be shown in the next section.

### 4.3.3. On the Relation between Adaptive Walks and Local Maxima

For the behavior of adaptive walks it is crucial how local maxima are distributed over the landscape and which shape they have. As a first order approximation, one can relate the walk length simply to the total number $N_{\max}$ or density $P_{\max}$ of maxima. In order to obtain a length scale $\lambda_{\max}$ related to the mean distance between local maxima, note that $1/P_{\max}$ is a "volume" associated with each maximum. One can then define $\lambda_{\max}$ implicitly by

$$\frac{1}{P_{\max}} = \sum_{d=0}^{\lambda_{\max}} \binom{L}{d}, \tag{4.79}$$

where the right-hand side is the number of genotypes that have distance $d \leq \lambda_{\max}$ to a given genotype, i.e., it is the volume of a ball with radius $\lambda_{\max}$. Its logarithm is approximately given by $\lambda_{\max} \cdot \log L$ and hence $\lambda_{\max} \approx -\log(P_{\max})/\log(L)$. As shown

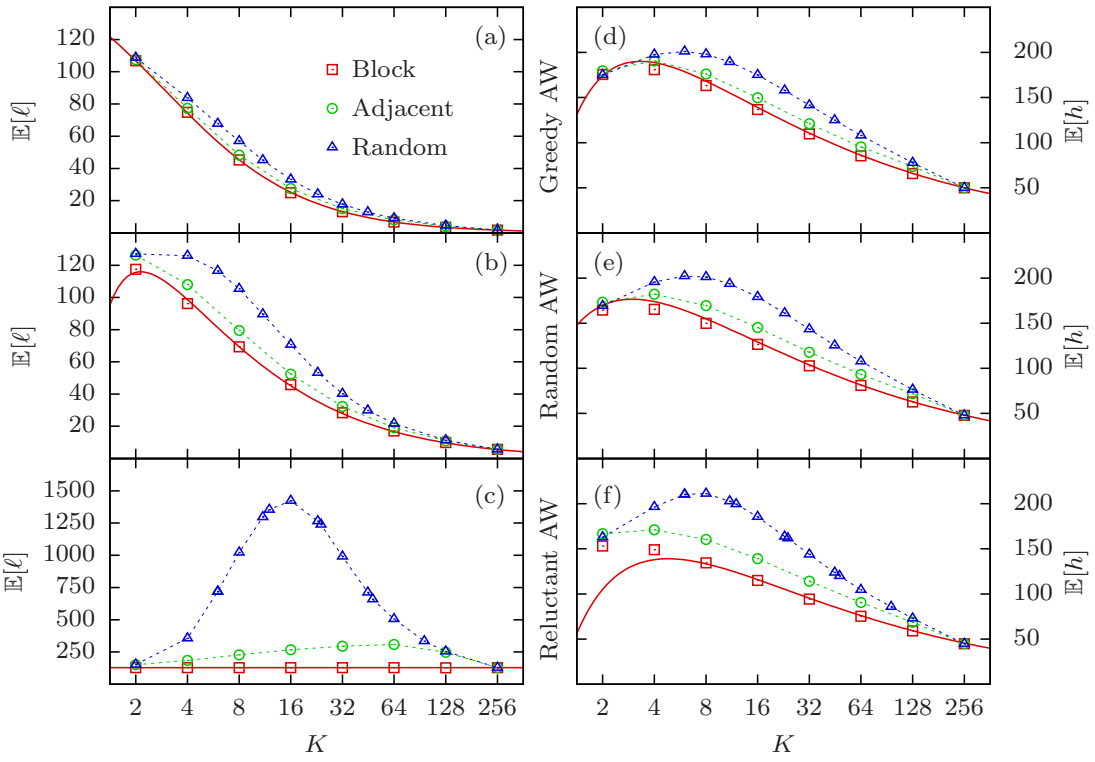



**Figure 4.5.** Walk length and height for three adaptive walk types and interaction schemes on an NK landscape with fixed $L = 256$. Symbols correspond to numerical results and are defined in panel (a), dashed lines are for visual guidance and solid lines are the results for the block model given by equations (4.76) and (4.77).

in figure 4.6(a), the mean length of greedy walks is proportional to $\lambda_{\max}$. Such a simple relation suggests that a greedy walker goes relatively straight to one of the nearest maxima. Conversely, since the other walk types do not show a linear behavior, their paths are more complicated and not only influenced by the maxima density but also by more subtle properties of the landscape.

This can also be seen by a more careful examination of walk heights. It was shown before that there are quantitative differences in adaptive walk heights on landscapes with different interaction patterns, but there are also qualitative differences. On the HoC landscape, the height increases with the greed of the walk type, i.e., greedy walks reach higher fitness than random AWs which in turn reach higher fitness than reluctant walks. According to equation (4.77), this behavior is inherited by the NK model with blockwise interactions. However, for random interactions and suitable values of $K$, the order is completely reversed. Reluctant walks become the most successful ones in terms of fitness, as can be seen in figure 4.7(a). This phenomenon was already observed before for similar landscape models in the context of, for instance, spin glasses [71] and complexity theory [74]. The effect is very counter-intuitive since it means that the highest fitness

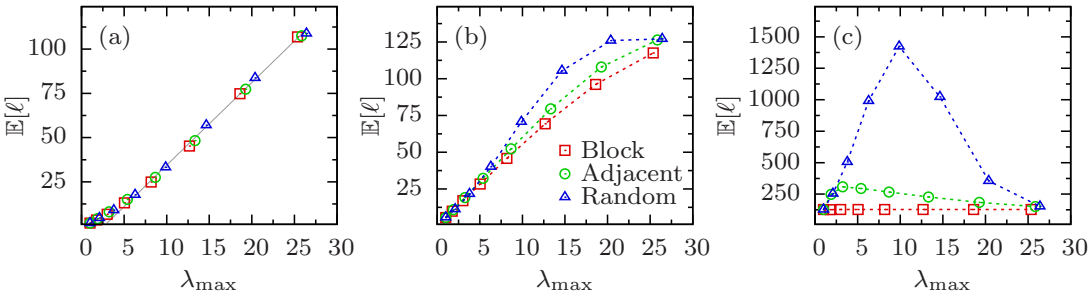



**Figure 4.6.** (a) Greedy, (b) random and (c) reluctant adaptive walk lengths as a function of the "length scale" $\lambda_{\max} = -\log(P_{\max})/\log(L)$ for an NK landscape with $L = 256$ and $K = 2^1, 2^2, \ldots, 2^8$. $P_{\max}$ is obtained by the numerical evaluation of equation (2.38). Different symbols correspond to different interaction schemes and are defined in panel (b). Dashed lines are for visual guidance, the solid grey line in panel (a) corresponds to a linear regression.

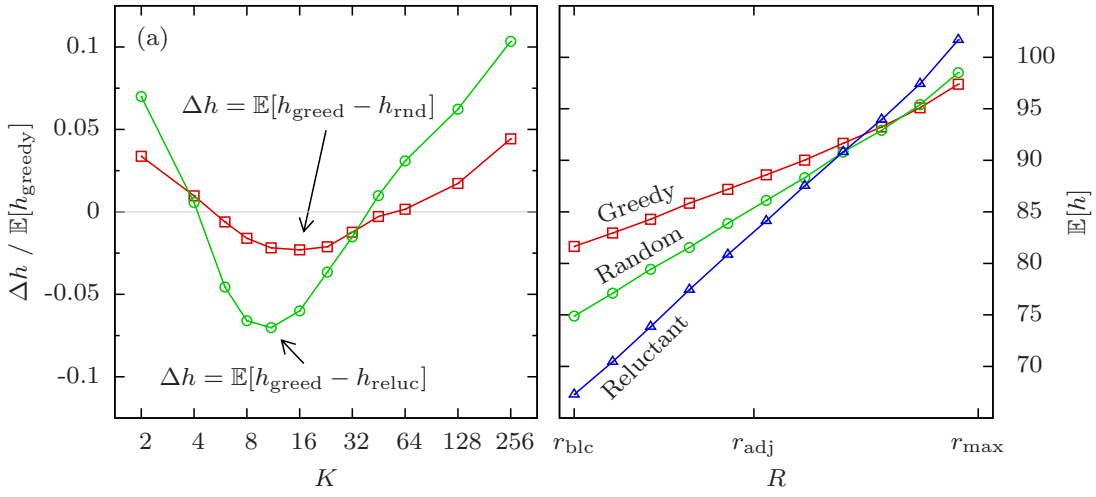



**Figure 4.7.** (a) Average height difference on an NK landscape with fixed $L = 256$ and random interaction scheme. Symbols correspond to numerical results and solid lines are for visual guidance. (b) Average height for constant $L = 128$ and $K = 8$ for varying rank $R$ of the interaction scheme. Lines are for visual guidance. Interaction schemes are produced by the procedure described in appendix A.2.1.

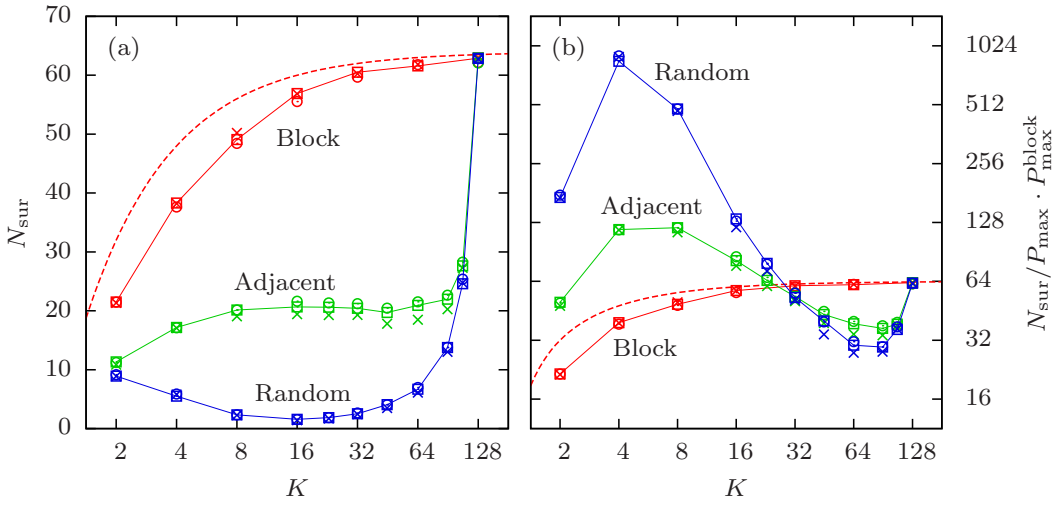



**Figure 4.8.** (a) Number $N_{\mathrm{sur}}$ of local maxima surrounding the final genotype of an adaptive walk. The underlying landscape has dimension $L = 128$. Different symbol shapes correspond to different walk types, but the influence of the latter is very small anyway and therefore hard to see. The dashed red lines shows $N_{\mathrm{sur}}$ for a randomly chosen maxima in the block model according to equation (2.41). (b) The same data, but scaled with the factor $P_{\mathrm{max}}^{\mathrm{block}}/P_{\mathrm{max}}$ by which the block model has more maxima. $P_{\mathrm{max}}$ is obtained from the numerical evaluation of equation (2.38). In both panels, solid lines are for visual guidance.

might be reached eventually if the walker always goes to the smallest possible fitness. Its impact on applications like optimization are probably rather small though, because the computation time of a reluctant walk is very high due to its enormous length compared to greedy walks. Therefore, if one is really interested in finding a particularly fit maximum, it would be still more efficient to perform several greedy walks rather than one reluctant walk.

A nice description of how the effect is influenced by the interaction pattern can be made in terms of the rank. The dependence of $h$ on $R$ for the different walk types is shown in figure 4.7(b). As one can see, the dependence is almost linear. While the $y$-intercept increases with greed of the walk type, the slope decreases. Therefore, depending on the parameters $L$ and $K$, there might be a point of intersection between the curves that marks the rank where random and reluctant walks, respectively, become more successful in reaching large fitness.

Why the success of adaptive walks depends on the fitness landscape in this way is still an open question. As a first hint one might ask how it looks like in the proximity of the genotype where the adaptive walk terminates. In the following, the quantity of interest is the (average) number $N_{\mathrm{sur}}$ of local maxima at distance $d = 2$ to that genotype, which is shown in figure 4.8. If local maxima were uncorrelated and unbiasedly found by adaptive walks, $N_{\mathrm{sur}}$ would be equal to $P_{\mathrm{max}}$ times the number $L(L-1)/2$ of genotypes at distance $d = 2$, but neither of these conditions is fulfilled as shown before. For blockwise interactions, $N_{\mathrm{sur}}$ can be calculated from equation (2.41) for a randomly

chosen maximum and is given by

$$N_{\mathrm{sur}}^{\mathrm{block}} = \frac{L(L-1)}{2}\frac{P_{\mathrm{max},2}}{P_{\mathrm{max}}} = \frac{L\,(K-1)}{2\,K}\,. \tag{4.80}$$

As can be seen in figure 4.8(a), $N_{\mathrm{sur}}$ is slightly smaller for maxima that are found by an adaptive walk in this case. For other interaction patterns, this quantity is not known for unbiased maxima, but maxima found in adaptive walks have much less other maxima nearby compared to the block model. This is not surprising since these interaction patterns produce also less maxima than the block pattern in total. Therefore, one should relate $N_{\mathrm{sur}}$ to the total number of maxima, which is increased in the block model by a factor of $\left(P_{\mathrm{max}}^{\mathrm{block}}/P_{\mathrm{max}}\right)$. If $N_{\mathrm{sur}}$ is divided by this value as in figure 4.8(b), it reveals that the clustering of maxima is actually smaller for block interactions and sufficiently small values of $K$. Unfortunately, one can not tell whether this is generic for local maxima in the NK model or only for those found by adaptive walks. The fact that $N_{\mathrm{sur}}$ is hardly influenced by the walk type might be a hint that also typical maxima show this phenomenon to some extent.

# 5. Recombination and Disruptive Selection

## 5.1. Wright Fisher Model

### 5.1.1. Classical Wright-Fisher Model

The Wright-Fisher model is an individual based model for population dynamics with discrete, non-overlapping generations. In each generation, all individuals are replaced by their offspring where the mean number of offspring corresponds to the fitness, i.e., the $i$-th individual with fitness $w_i$ will have offspring drawn from some distribution with mean value $w_i$, e.g, a Poisson distribution. Let $m_i(t)$ denote the number of offspring that the $i$-th individual in generation $t$ will leave to the next generation. Then in generation $t+1$ there will be $M_{t+1} = \sum_i m_i(t)$ individuals. Assuming that the $m_i$ are Poisson distributed, also $M_{t+1}$ will be Poisson distributed with

$$\mathbb{P}[M_{t+1} = N \mid M_t = M] = \frac{(M\bar{w}_t)^N}{N!} e^{-M\bar{w}_t}\,, \tag{5.1}$$

where

$$\bar{w}_t = \frac{1}{M_t} \sum_{i=1}^{M_t} w_i(t) \tag{5.2}$$

is the average fitness in generation $t$. Since individuals will produce offspring independently, the joint distribution for the individual offspring is given by the product

$$\mathbb{P}[m_1(t) = n_1, \ldots, m_M(t) = n_M \mid M_t = N] = \prod_{i=1}^{N} \frac{w_i^{n_i}}{n_i!} e^{-w_i} = e^{-N\bar{w}_t} \prod_{i=1}^{N} \frac{w_i^{n_i}}{n_i!}\,. \tag{5.3}$$

The actual Wright-Fisher model [13, 14] now arises from the claim that the population size is constantly $N$, i.e., the population's distribution in the next generation is given by

$$\begin{aligned}
&\mathbb{P}[m_1(t) = n_1, \ldots, m_N(t) = n_N \mid M_t = M_{t+1} = N] \\
&= \frac{\mathbb{P}[m_1(t) = n_1, \ldots, m_N(t) = n_N \mid M_t = N]}{\mathbb{P}[M_{t+1} = N \mid M_t = N]} \\
&= \frac{N!}{\prod_{i=1}^{N} n_i!} \prod_{j=1}^{N} \left( \frac{w_i}{N\bar{w}_t} \right)^{n_i},
\end{aligned} \tag{5.4}$$

if $\sum_i n_i = N$ and zero otherwise. This is a multinomial distribution with probabilities $w_i/N\bar{w}_t$. Therefore, a common interpretation of the dynamics is that individuals in

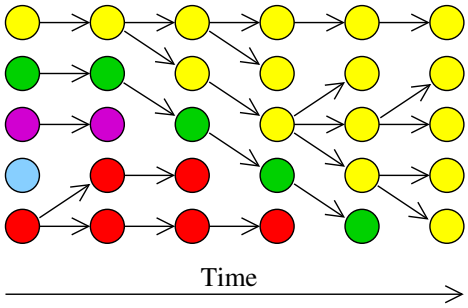



**Figure 5.1.** Illustration of the original Wright-Fisher dynamics with $N = 5$ individuals over 6 generations. Colors indicate the ancestor from the first generation. Each arrow goes from a parent to its child.

generation $t+1$ randomly "choose" their parent to be the $i$-th individual from generation $t$ with probability $w_i/N\bar{w}_t$. An illustration of this process can be found in figure 5.1.

In this form, the model does not include any notion of genotype space or mutations, but they can be easily included by assigning a genotype $\sigma$ to each individual. The individual's fitness $w_i$ is then given by the fitness $w(\sigma)$ of its genotype. A child will carry the same genotype as its parent, unless a mutation occurs with some probability $\mu$. The variation of the standard Wright-Fisher model including these mechanisms, which is used in this thesis, will be explained in the next section.

### 5.1.2. Wright-Fisher Model on the Hypercube

As mentioned above, a version of the Wright-Fisher model will be used which resembles the original dynamics but is not equivalent. The reason for this is simply to speed up the computation of the simulations. On the hypercube, it is more convenient to formulate the model in terms of occupation numbers of genotypes rather than individuals. This is because individuals carrying the same genotype are identical, at least within the frame of the Wright-Fisher dynamics. Let $N_\sigma$ be the number of individuals carrying genotype $\sigma$ and

$$\bar{w} = \frac{1}{N} \sum_{\sigma} w(\sigma)\, N_\sigma \tag{5.5}$$

the fitness averaged over the population. The dynamics is then composed of single steps that alter the occupation numbers. Mutation and selection are actually stochastic processes, but their corresponding steps are treated deterministically here. Note that this leads to non-integer values of the occupation numbers $N_\sigma$, but the effect is negligible for large $N$. Stochasticity is included afterwards by an explicit random sampling step. The definition of each step is given in the following.

**Mutation.** A fraction $\mu$ of all individuals mutate. All mutants are evenly distributed to genotypes that are neighbors of their parents genotype. In formulas:

$$N_\sigma \to (1-\mu)\, N_\sigma + \frac{\mu}{L} \sum_{\tau \in \mathcal{U}_\sigma} N_\tau \,.$$

**Selection.** The number of individuals carrying a certain genotype is updated according to the relative fitness of that genotype, i.e.,

$$N_\sigma \to \frac{w(\sigma)}{\bar{w}} N_\sigma \,.$$

**Random Sampling.** All occupation numbers $N_\sigma$ are replaced by a random number drawn from a Poisson distribution with parameter $N_\sigma$:

$$N_\sigma \to \mathrm{Poi}(N_\sigma) \,.$$

Actually, one should use a multinomial distribution in order to resemble equation (5.4). However, performing the sampling independently for each genotype saves a lot of computation time when $L$ is large [83].

**Normalization.** Since this way of random sampling does not guarantee the total number of individuals to be equal to $N$, one has to perform the normalization explicitly:

$$N_\sigma \to \frac{N}{\sum_\tau N_\tau} N_\sigma \,.$$

Note that one can easily include further mechanisms as additional steps or alter existing ones due to the modular structure of this model. In section 5.2, an additional step "Recombination", will be added to this dynamics. In section 5.3, the selection step will be altered in order to allow for competition between individuals.

### 5.1.3. Observables and General Behavior

The most natural quantity to measure the success of a population as a whole is the fitness $\bar{w}$ averaged over the population as defined in equation (5.5). It will be presented in the following mostly as a time series, i.e., in dependence on the generation.

Another useful quantity will be the *diversity* defined as

$$D = \exp\left[ -\sum_\sigma n_\sigma \log(n_\sigma) \right], \tag{5.6}$$

where $n_\sigma = N_\sigma / N$ is the fraction of the population carrying genotype $\sigma$. As one can see, it is the exponential of the population's entropy inspired by the following intuition: Suppose there exist $n$ different genotypes and all of them are carried by the same amount of individuals, i.e., $n_\sigma = 1/n$. In this case, $D$ is just the number of occupied genotypes. However, the population is usually not uniformly distributed which will cause a decrease of the diversity. This is intentional since it causes that $D$ is close to 1 if only one genotype is macroscopically occupied, even though it is surrounded by several mutants. In some

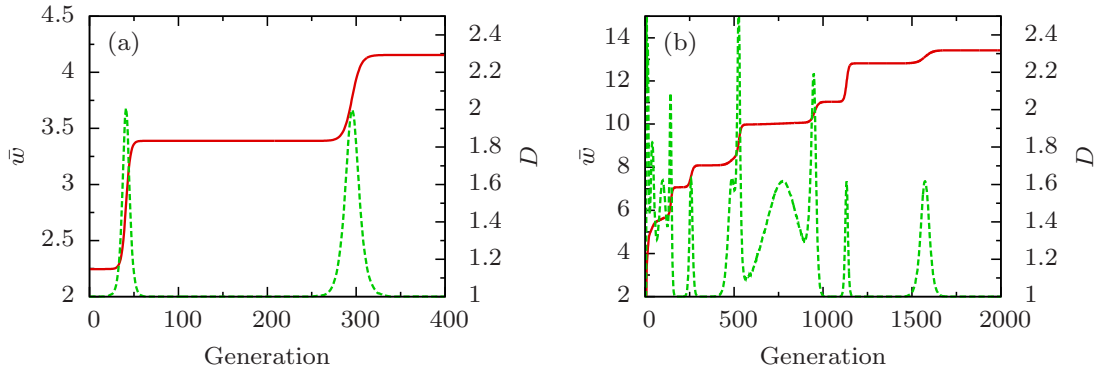



**Figure 5.2.** Example of time series for mean fitness $\bar{w}$ (solid lines) and diversity $D$ (dashed lines) of a population with $N = 10^5$ and $N\mu = 1$. The underlying fitness landscape is a standard exponentially distributed HoC landscape in panel (a) and a RMF landscape with slope $s = 1$ and standard exponentially distributed noise term in panel (b). Initially, the whole population is located at the reference sequence $\tilde{\sigma}$, which corresponds to a randomly chosen genotype in the HoC case and to a poorly adapted state in the RMF case.

sense, the diversity can be thought of as the effective number of occupied genotypes without the need of defining a threshold at which a genotype is considered as occupied.

Examples for time series of $\bar{w}$ and $D$ are shown in figure 5.2 for populations evolving on a HoC and on a RMF landscape. They are supposed to reflect the typical behavior. As one can see, the mean fitness increases monotonically but stepwise. Each step corresponds to the population moving to a new genotype with larger fitness. As described earlier, this moving process begins with the rise of a mutant carrying the new genotype. Since it has larger fitness than the prevalent genotype, there is a positive fixation probability. Note that fixation is not meant in a strict sense here, it rather refers to states where the population is highly concentrated on a genotype. Since the fixation process happens gradually, there are a few generations where the old and new genotype are both macroscopically occupied, which appears as a peak in the diversity $D$. Obviously, this adaptation process happens faster on the less rugged RMF landscape than on the HoC landscape. Note that, even if $N\mu$ is of order 1, the time between two fixation events is still much larger than the fixation time. The diversity is very close to 1 between these events and hence the population can still be thought of as a single entity traveling through the genotype space, i.e., similar to an adaptive walk. An important difference to adaptive walks is that the dynamics does not end when the population gets trapped, i.e., when it fixates on a local maximum. This delays the adaptation process by quite a lot, but the population can escape eventually.

## 5.2. Recombination

### 5.2.1. Advantages and Disadvantages of Recombination

Recombination is the creation of novel genetic information of an individual, the child, from the genetic information of parental individuals. Findings in paleontology suggest that recombination exists as early as one billion years ago [84] and currently it is prevalent in nature, at least among higher organisms, indicating an advantage under various circumstances. This advantage, however, is far from being obvious [85–88]. Sexual reproduction in particular seems grossly inefficient since all individuals of a population need about the same amount of resources while only females give birth to offspring, an issue known as the *two-fold cost of sex* [89, 90]. But even when this effect and other implications arising from two sexes are ignored, there are still disadvantageous effects of recombination. A well-known example is the *recombination load* [91, 92], i.e., the issue that reshuffling a genotype can result in the disruption of beneficial combinations of genes.

Nevertheless, there are also lots of positive effects proposed in the literature. The *Weismann effect* [93, 94], for instance, refers to the fact that the amount of genetic variability is increased by recombination. This facilitates the search for particularly fit genotypes but can also be disadvantageous when an optimal genotype was already found. Probably the most noted advantage is *Muller's ratchet* [95–97] which states that asexual populations accumulate deleterious mutations almost irreversibly since they can only be purged if a back-mutation occurs, which is very unlikely for large $L$. Recombination, on the other hand, gets rid of those mutations easily. Conversely, if two different beneficial mutations arise in two different individuals of an asexual population, they will compete for fixation. At best, this delays the emergence of an individual carrying both mutations while recombination obviously facilitates it, a scenario known as the *Fisher-Muller* or *Hill-Robertson effect* [13, 96, 98, 99]. However, note that both Muller's ratchet and the Hill-Robertson effect have limited validity on a sign-epistatic landscape since the effect of mutations depends on the genetic background. Deleterious mutations can turn into beneficial ones when another mutation occurs and the combination of two beneficial mutations can be deleterious. Epistasis and local maxima in particular obstruct adaptation of sexual populations. A theoretical study on a double-peaked two-locus landscape showed that high rates of recombination inhibit substantially the escape from the smaller maximum to the larger one, even though the process is accelerated slightly for small recombination rates [100]. This *trapping* at local maxima will turn out to be an important factor on multidimensional landscapes.

Another important concept is the *Red Queen hypothesis* [101]. It states that organisms need to adapt constantly to an ever-changing environment. The advantage of recombination is due to a faster response to environmental changes rather than the ability to find particularly fit states eventually. The name of this concept is derived from Lewis Carroll's famous book *Through the Looking-Glass* [102] where the Red Queen explains to Alice that "it takes all the running you can do, to keep in the same place."

### 5.2.2. Wright-Fisher Dynamics with Recombination

How the new genotype is composed of parts from the parents genotype can generally be very complex, but here only a simple variant, the *uniform crossover*, will be studied: Each locus is either taken from the first or the second parent with equal probability. In the optimized Wright-Fisher model that will be used, the whole population is replaced by new individuals whose genotype is the recombination of two randomly chosen parents. This additional step is included right after the normalization step. The focus will be on the comparison of non-recombining and recombining populations, also referred to as asexual and sexual populations, respectively, in the text. Despite this naming, the setting corresponds to mere genetic recombination and not to actual sexual reproduction since distinct genders of individuals and consequences thereof, e.g., mating, will not be considered.

The underlying fitness landscape will be the RMF landscape with a random part drawn from an exponential distribution. The distribution's mean value $\lambda$ will be the main parameter to control the landscape ruggedness. Unless otherwise stated, all individuals carry initially the reference genotype $\tilde{\sigma}$ of the RMF model, i.e., the scenario correspond to a situation where the population is poorly adapted in the beginning. Simulation results will be mostly presented as averaged time series of the fitness difference $\Delta w = \langle w_{\mathrm{rec}} \rangle - \langle w_{\mathrm{norec}} \rangle$ between sexual and asexual populations. An advantage of recombination corresponds to $\Delta w > 0$. The angle brackets $\langle w \rangle$ denote here and in the following the average of $\bar{w}$ over realizations, i.e., it includes two averages: One over the population and one over realizations.

The scenario might seem a bit restrictive, but several variations were tested and indicate that qualitative results are quite robust [103]. For instance, the NK model, with $K$ playing the role of a ruggedness parameter, may also serve as landscape. With regard to the dynamics, variants including an infinite population size or direct competition between recombining and non-recombining individuals show qualitatively the same behavior as well.

### 5.2.3. When is Recombination Beneficial?

To begin with, the averaged time series of of the mean fitness $\langle w \rangle$ of both sexual and asexual population are shown in figure 5.3(a). As expected, $\langle w \rangle$ increases with increasing time and ruggedness parameter $\lambda$. The differences $\Delta w$ between recombining and non-recombining populations can be best seen in figure 5.3(b). The typical shape of $\Delta w$ includes a minimum after a few generations which is followed by an ascent leading to a maximum and finally a decline. A close look at the curve for $\lambda = 3$ reveals that there is also a second minimum and maximum, even though it is less pronounced. As one can see, the answer to the question whether recombination is advantageous or not depends on the time when it is asked. At very small and large times, recombination seems to be always the worse option, but in between there might be a time window where $\Delta w$ is positive. However, even if a maximum of $\Delta w$ exists, it might have a value below zero, as in the curve for $\lambda = 3$. Generally, the advantage of recombination, whether measured

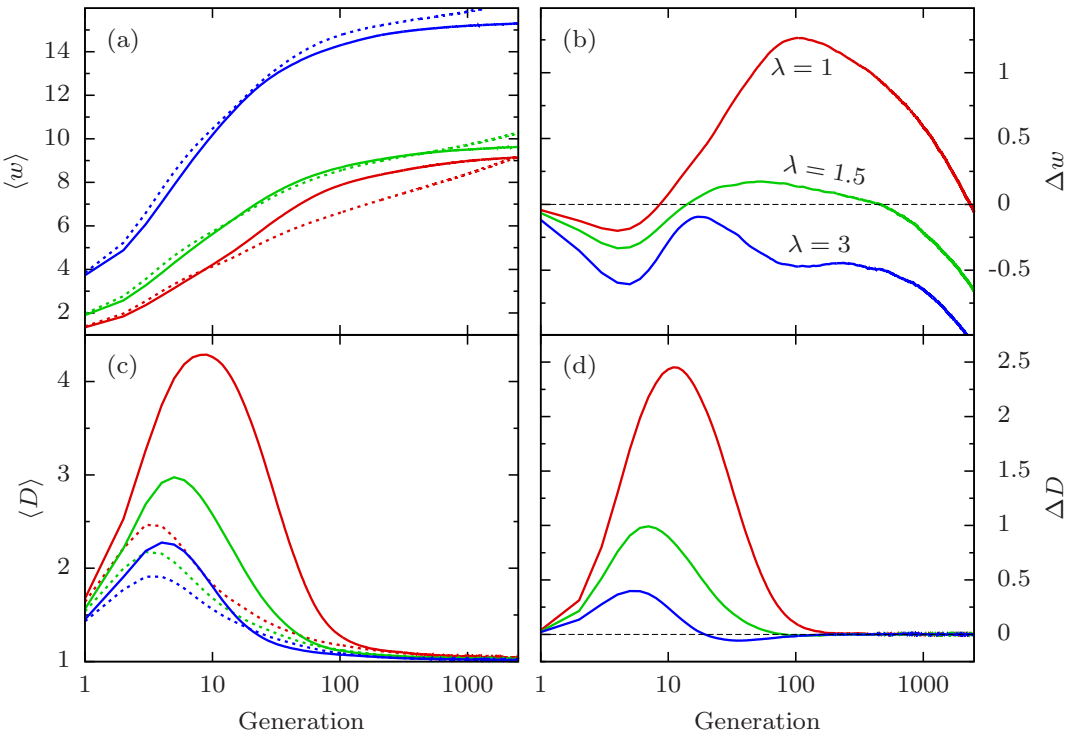



**Figure 5.3.** (a) Fitness time series of recombining (solid lines) and non-recombining (dashed lines) populations. (b) Time series of fitness difference $\Delta w$. (c) Diversity time series of recombining (solid lines) and non-recombining (dashed lines) populations. (d) Time series of diversity difference $\Delta D$ between recombining and non-recombining populations. In all panels, different colors correspond to different noise strengths $\lambda$ and are defined in panel (b). The other parameters are given by $L = 16$, $s = 1$, $N = 1000$, and $N\mu = 2$.

by its duration or its maximal fitness advantage, appears to decreases with increasing ruggedness of the landscape.

At the same time, sexual and asexual populations have very different behavior with regard to their diversity $D$. As shown in figure 5.3(c) and (d), sexual populations have substantially larger diversity than asexual ones, at least in the beginning. This is caused by the fact that the initial genotype is a poorly adapted state and hence almost all mutations will lead to fitter states. These mutants will co-exist until the fittest of them will be selected. Since the fitness values are more similar if the noise term controlled by $\lambda$ is weaker, selection is weaker as well and hence the diversity increases with decreasing $\lambda$. This happens for both recombining and non-recombining populations, but during the phase of many co-existing mutants, recombination will produce lots of offspring with novel genotypes. In other words, recombination amplifies the already existing diversity in the initial phase of the dynamics, in accordance with the Weismann effect. After some

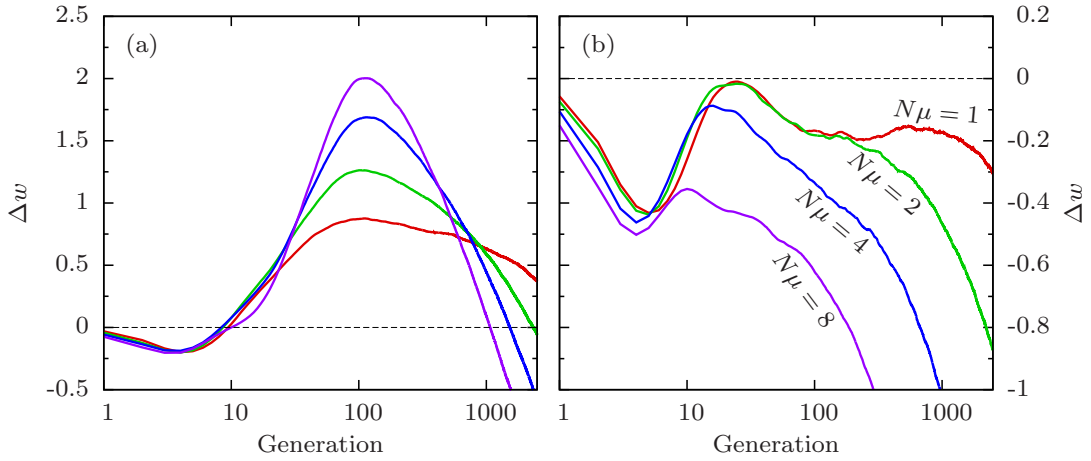



**Figure 5.4.** Time series of fitness difference $\Delta w$ for a landscape with (a) $\lambda = 1$ and (b) $\lambda = 2$. Different colors correspond to different values of $\mu$ and are defined in panel (b). The other parameters are given by $L = 16$, $s = 1$, and $N = 1000$.

time, $\langle D \rangle$ levels off for both population types to a value slightly above 1 which indicates that they become quasi-monomorphic.

Different values of $N\mu$ control the supply of mutants. As recombination needs a sufficient amount of mutants to be effective, one might guess that the effect of recombination is intensified if $N\mu$ gets larger. This is mostly true, but note that it does not necessarily mean that recombination becomes more advantageous since it has several downsides, too. If the landscape is too rugged, increasing $N\mu$ is disadvantageous for sexual populations, while it is advantageous on less rugged landscapes. This can be seen from the comparison of figures 5.4(a) and (b). The influence of ruggedness on the success of recombination will be discussed in the next section in more detail.

### 5.2.4. Temporal Pattern of $\Delta w$ and Dynamic Regimes

For the study of the temporal pattern of $\Delta w$, one should go back for a moment to the individual fitness curves $\langle w_\mathrm{r} \rangle$ and $\langle w_\mathrm{nr} \rangle$ for recombining and non-recombining populations, respectively, that are shown in figure 5.3(a). While the slope of $\langle w_\mathrm{r} \rangle$ is larger than $\langle w_\mathrm{nr} \rangle$ for intermediate times, it is smaller in the very beginning and the end. This divides the temporal pattern roughly in three regimes: In the initial and final regime, $\langle w_\mathrm{nr} \rangle$ grows faster than $\langle w_\mathrm{r} \rangle$, in the intermediate regime it is the other way round.

The reason why $\langle w_\mathrm{nr} \rangle$ grows faster than $\langle w_\mathrm{r} \rangle$ in the initial regime can be ascribed mainly to the Weismann effect, i.e., the diversity $D$ of recombining populations is so large that also many sites on the landscape are populated that are not particularly fit. Even though $\Delta w$ is negative in this regime, it can actually be seen as an prospective advantage since the fittest individual among both populations belongs to the recombining population with high probability, but beside many less fit individuals. However, the

adaptation of a population as a whole is mostly driven by the fittest individual and therefore fitness gain will be even accelerated once the diversity is reduced when less fit mutants become extinct.

Diversity reduction also marks the transition to the intermediate phase. This dynamic regime is characterized by smaller diversity and comparatively fast adaptation, indicating that fitter genotypes get populated sequentially. Given that $N\mu$ is sufficiently large such that there is a steadily supply of new mutants, one can ascribe the fact that $\langle w_\mathrm{r} \rangle$ grows faster than $\langle w_\mathrm{nr} \rangle$ in this phase largely to the Hill-Robertson effect. It is particularly effective on an RMF landscape due to its inherent fitness gradient. If two mutants are recombined that are both located one step uphill with respect to the most populated state, the child individual will be located two steps uphill with a certain probability. Such events accelerate the adaptation process substantially. Of course, recombination can also produce individuals that are located downhill or have lower fitness due to the noise term, but they will be purged quickly by selection.

Eventually, this sequential dynamics ends when the population reaches a local maximum. Further adaptation can then only happen if a fitness valley is crossed. As mentioned earlier, it was found on a two-locus landscape that escaping from local maxima is very difficult for recombining populations [100]. In order to address the question of whether this is relevant on a high-dimensional landscape, escape and trapping events will be counted in the following. The population is regarded as trapped if 70% of all individuals are located on the same local maximum. An escape event is registered if this fraction drops below 50% for a formerly trapped population. The fraction $p_\mathrm{esc}$ of populations that escaped from the first maximum they get trapped on before the 2500-th generation, which is the last generation in figures 5.3 and 5.4, is shown in figure 5.5(a). As expected, $p_\mathrm{esc}$ decreases with increasing $\lambda$ and decreasing mutation rate $\mu$ since local maxima become more peaked and a larger amount of mutants is produced to pull the population from the maximum, respectively. More importantly, the number of escape events drops substantially to a tiny fraction if recombination is active. To some extent, trapping is delayed for sexual populations due to their larger diversity as can be seen in figure 5.5(b), but the effect is much weaker than the inability to escape.

If at all possible, recombining populations build up their advantage in the intermediate regime. The more rugged a landscape is, the more difficult is it to escape from maxima, the earlier populations get trapped and, in turn, the shorter is the time of the intermediate regime. Asexual populations can continue adapting after trapping, even if they are slowed down quite a lot, but appreciable dynamics of recombining populations basically stops in most realizations as soon as a local maximum is reached. As a consequence, recombination is more advantageous if the landscape is less rugged. With these finding, one can also explain the different response to an increase of the mutation rate in figure 5.4. On the less rugged landscape ($\lambda = 1$), an increased supply of mutants simply amplifies the advantage that recombining populations can build up in the intermediate regime. They get also trapped earlier such that $\Delta w$ is lower in the long run and also the time of the advantage is shorter, but the peak of the advantage increases. On the rugged landscape with $\lambda = 2$, populations get trapped earlier in the simulation, in particular before recombining populations can build up the advantage.

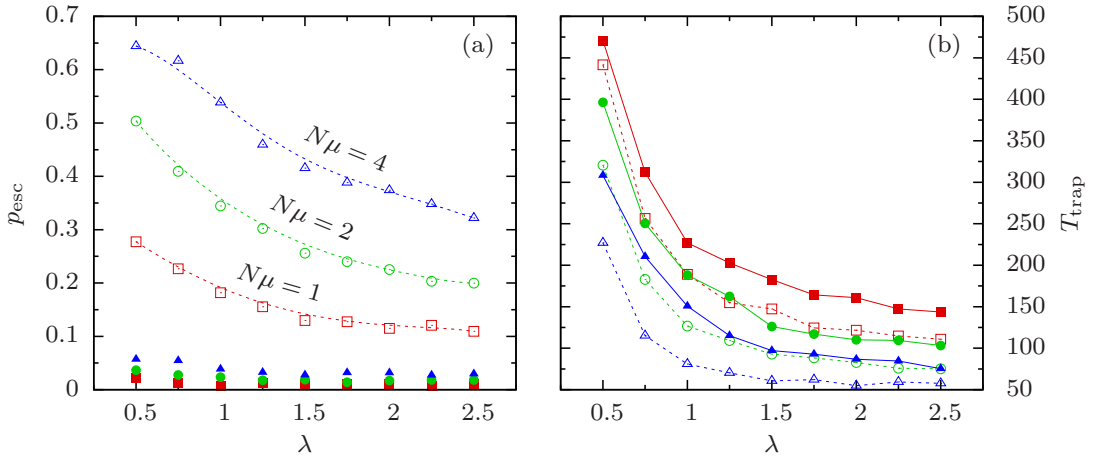



**Figure 5.5.** (a) Fraction $p_{\mathrm{esc}}$ of populations that escaped from the first maximum they get trapped on. Open symbols correspond to asexual, filled symbols to sexual populations. Different colors and symbol shapes are to distinguish different mutation rates. (b) Mean time $T_{\mathrm{trap}}$ at which populations get trapped on a local maximum for the first time. Symbols and colors have the same meaning as in panel (a). In both panels, the parameters are $L = 16$, $s = 1$, and $N = 1000$. Lines are for visual guidance, solid ones for recombining, dashed ones for non-recombining populations.

The latter does not change much if $\mu$ is increased, but trapping occurs even earlier. Asexual populations, on the other hand, benefit much more from an increase of $N\mu$ and hence $\Delta w$ decreases at almost all times.

### 5.2.5. Mechanisms to Prolong the Advantage of Recombination

As it was shown, there is a transitory advantage of sex on landscapes with sufficiently small ruggedness, but there are several ways to enhance this advantage further. One hurdle for recombination is that the models used here drive the population to an almost monomorphic state. As will be discussed later in section 5.3, the version of the Wright-Fisher model that is used here does not allow for stable coexistence of subpopulations carrying different genotypes. Therefore, recombination acts mostly on pairs of individuals that have a rather short genetic distance. Disruptive selection [104], a mechanism that enables or even expedites the parallel existence of distant populated genotypes, could benefit recombining populations quite a lot. It is not only that recombination of distant genotypes renders the exploration of large parts of the genotype space possible, a population that is spread out is also prevented from trapping on local maxima to some extent.

Another option to enhance the success of recombination would be to introduce recombination schemes that take the genetic structure into account. Take for instance the NK model with blockwise interaction: Epistasis only exists within the fixed blocks. If recombination works in a way that preserves the blocks, the mechanism that causes strong trapping would not work since the landscape is then additive with respect to

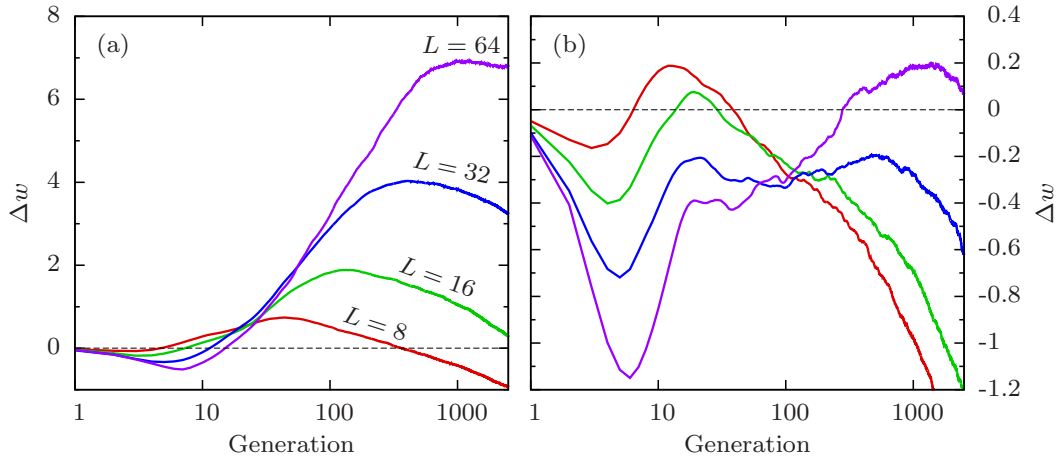



**Figure 5.6.** Time series of fitness difference $\Delta w$ for a landscape with (a) $\lambda = 1$ and (b) $\lambda = 2$. Different colors correspond to different dimensionality $L$ and are defined in panel (a). Other parameters are $N = 2000$, $N\mu = 4$, and $s = 1$.

recombination. More on the relation between genetic structure and recombination can be found in [105].

Also an increase of the genome size $L$ can strengthen the success of recombination. So far, only a rather short sequence length of $L = 16$ was studied. The influence of $L$ on $\Delta w$ can be found in figure 5.6. Increasing $L$ increases the distance that the population can travel through the genotype space before it gets trapped on a local maximum. Therefore, one should expect a prolonged duration of the intermediate regime, but the duration and adaptation speed of the other dynamic regimes are affected as well. This results in a complicated dependence on the genome size, similar to the previously shown alteration of the mutation rate $\mu$. On a landscape with small ruggedness, as it is shown in panel (a), the behavior is still quite simple as an increased dimensionality seems to alter the duration and effect of the dynamic regimes evenly such that the curves of $\Delta w$ stay qualitatively the same. As a consequence, one can clearly see that the advantage of recombination increases with increasing $L$. For the more rugged landscape shown in panel (b), the situation is much more complicated. The shape of $\Delta w$ changes in a way that a second maximum arises while the first one vanishes when $L$ is increased. Interestingly, this results in a non-monotonic behavior of $\Delta w$'s maximal value, i.e., an advantageous time window of recombination can be found for $L = 16$ and $L = 64$ but not for $L = 32$.

Finally, there is also a way to maintain the advantage of recombination indefinitely: Fitness seascapes, which will be discussed in the next section in more detail.

### 5.2.6. Stationary Advantage on Fitness Seascapes

As shown, recombining populations lose the evolutionary race against non-recombining populations on rugged fitness landscapes in the long run. There might be an

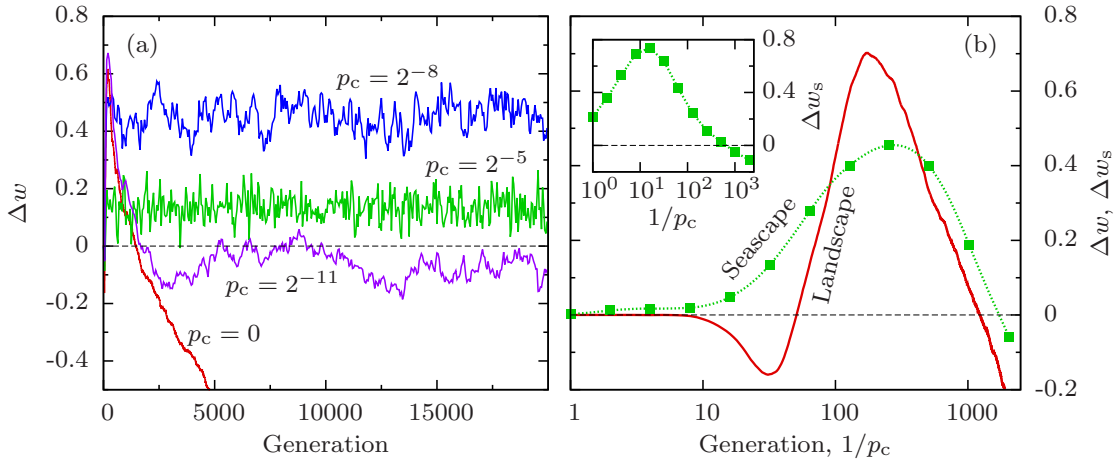



**Figure 5.7.** (a) Time series of $\Delta w$ on a seascape with hard reset. Parameters are $L = 16$, $s = 1$, $\lambda = 0.75$, $N = 2000$, and $N\mu = 5$. (b) Symbols show the stationary value of panel (a), averaged from generation $10^4$ to $2 \cdot 10^4$, in dependence on the mean time $1/p_c$ between two resets. For comparison, the red solid line shows the time series of $\Delta w$ for a population starting on a randomly chosen genotype of an ordinary landscape with the same parameters. The inset shows the stationary advantage on a seascape with soft reset. Green dotted lines are for visual guidance.

advantageous period for recombination, that can even be prolonged by the mechanism described above, but the advantage is mostly transitory. However, note that in many scenarios considered here, the advantage lasts up to a few thousand generations. In such a long time, the circumstances under which the population evolves might change as a result of, for instance, a changing environment or a population traveling through a spatially inhomogeneous world.

This scenario can be modeled via a *fitness seascape*, i.e., a time-dependent fitness landscape. To keep things simple, the seascape used here will be modeled as an ordinary landscape with fitness values that are changed in irregular intervals. More precisely, changes will happen after each generation of the Wright-Fisher dynamics with probability $p_c$. Then the random component $\eta(\sigma)$ of all fitness values will be drawn anew and the reference sequence $\tilde{\sigma}$ will be set to another position of the hypercube. With regard to the latter, two versions will be distinguished: The new reference sequence is chosen randomly either from the neighbors of the old sequence (*soft reset*) or from all genotypes (*hard reset*). Time series of $\Delta w$ on a seascape with hard reset are shown in figure 5.7(a).

On a seascape, local maxima only exist temporarily. Since they are almost the only thing that prevents recombination from being superior, one can observe an advantage of sex even in the long run for suitable choices of $p_c$. After each reset, the population will start an adaptation process from anew, resulting in a stationary value of the fitness after some time for both sexual and asexual population and, in turn, also for the fitness difference $\Delta w$. The stationary value $\Delta w_s$ of the latter is shown in

figure 5.7(b) in dependence on the mean time $1/p_c$ between two resets of the landscape. As expected, $\Delta w_s$ is positive for both soft and hard reset and a rather large variety of values for $p_c$. Also not surprising is the fact that the dependence is non-monotic. For large $p_c$, the population's fitness is dominated by fluctuations of the landscape rather than its adaptational success and hence recombination does not make a big difference. Nevertheless, in case of the soft reset, even the extreme case $p_c = 1$ shows a significant advantage. If, on the other hand, $p_c$ is too small, the lifetime of local maxima becomes larger than the escape time of asexual populations and hence they will have the same advantage as on a static landscape.

The largest advantage arises when the time $1/p_c$ is of the same order as the time $t_{max}$ of the maximum on a static landscape with initial conditions corresponding to the situation right after a reset. While these conditions are hard to find for the soft reset, they are simply a population starting from a randomly chosen genotype in case of the hard reset. As one can see in 5.7(b), the stationary advantage on a time-dependent landscape behaves similar to the time-dependent advantage on a stationary landscape.

## 5.3. Frequency-Dependent and Disruptive Selection

In this section the Wright-Fisher model will be altered slightly such that explicit competition between individuals is included. One way to model this is to define an effective, frequency-dependent fitness

$$w_e(\sigma) = w(\sigma)\,[1 - \beta\, n_\sigma]\,, \tag{5.7}$$

where $\beta \in [0, 1]$ is a parameter that controls the strength of interactions. Obviously, the effective fitness of $\sigma$ becomes smaller the more individuals carry that genotype. The interpretation is that genetically identical individuals fight for exactly the same resources which may lead to a shortage. Interactions are zero-ranged by that definition, i.e., individuals carrying a genotype $\tau$ with $d(\sigma, \tau) = 1$ do not have influence on the effective fitness of individuals carrying genotype $\sigma$ and vice versa. This might not be very realistic, but it is straightforward to extend the definition of the effective fitness by also including further afar genotypes with an influence that decays with genetic distance. However, no qualitative differences to longer-ranged interactions were found for the scenarios considered here. For the sake of simplicity, zero-ranged interactions will be kept. In the following, consequences of the frequency-dependent selection will be outlined.

### 5.3.1. Coexistence

The Wright-Fisher model in the form described in section 5.1.2 does not allow for stable coexistence of sub-populations at different genotypes. Suppose two genotypes $\sigma$ and $\tau$ are occupied by a fraction $n_\sigma = n$ and $n_\tau = 1 - n$ of individuals, respectively. If

mutations and random fluctuations are neglected, the population evolves according to

$$n(t+1) = n(t)\frac{w(\sigma)}{\bar{w}} = \frac{n(t)\,w(\sigma)}{n(t)\,w(\sigma) + [1-n(t)]\,w(\tau)}\,. \tag{5.8}$$

This equation has only two fixed points, $n_0^* = 0$ and $n_1^* = 1$, that correspond to monomorphic populations. If $w(\sigma) > w(\tau)$, the fixed point $n_1^*$ is stable while $n_0^*$ is unstable. For $w(\sigma) < w(\tau)$, it is the other way round.

With frequency-dependent selection, the equation reads

$$n(t+1) = n(t)\frac{w_\mathrm{e}(\sigma)}{\bar{w}_\mathrm{e}} = g(n(t))\,, \tag{5.9}$$

where

$$g(n) = \frac{n\,w(\sigma)\,[1-\beta\,n]}{n\,w(\sigma)\,[1-\beta\,n] + (1-n)\,w(\tau)\,[1-\beta\,(1-n)]}\,. \tag{5.10}$$

One has still the monomorphic fixed points $n_0^* = 0$ and $n_1^* = 1$ with this equation, but an additional fixed point

$$n_2^* = \frac{w(\sigma) - w(\tau)\,(1-\beta)}{\beta\,[w(\sigma) + w(\tau)]} \tag{5.11}$$

arises that corresponds to coexistence of individuals carrying $\sigma$ and $\tau$. As expected, one has $n_2^* = 1/2$ in the neutral case $w(\sigma) = w(\tau)$. Concerning their stability, only one of the three fixed points is stable while the others are unstable: $n_0^*$ is stable if $w(\sigma) < w(\tau)\,(1-\beta)$, $n_1^*$ is stable if $w(\sigma) > w(\tau)/(1-\beta)$, and $n_2^*$ is stable if

$$w(\tau)\,(1-\beta) < w(\sigma) < w(\tau)/(1-\beta).$$

Note that the region where $n_2^*$ is unstable corresponds to unphysical values $n_2^* \notin [0,1]$. Put simply, coexistence is possible if the fitness differs by less than a factor of $(1-\beta)$. The overall fixed point structure is illustrated in figure 5.8.

### 5.3.2. Behavior on High-Dimensional Landscapes

As shown above, competition largely prevents the population to evolve into a monomorphic state. This means in particular that also trapping at local maxima becomes unlikely since it is only possible if the fitness of a maximum is more than $1/(1-\beta)$ times larger than that of its surrounding genotypes. Therefore, one can expect that the adaptation process is generally accelerated and that the effect becomes stronger with increased interaction strength $\beta$.

On the other hand, competition also has disadvantages for the population. For instance, if a particularly fit genotype $\sigma$ is found, a smaller fraction of the population will carry it because otherwise its effective fitness $w_\mathrm{e}(\sigma)$ will drop. More importantly, the relevant measure for the success of populations is the mean effective fitness $\bar{w}_\mathrm{e}$ rather than the native fitness $\bar{w}$. By definition, the effective fitness decreases with interaction

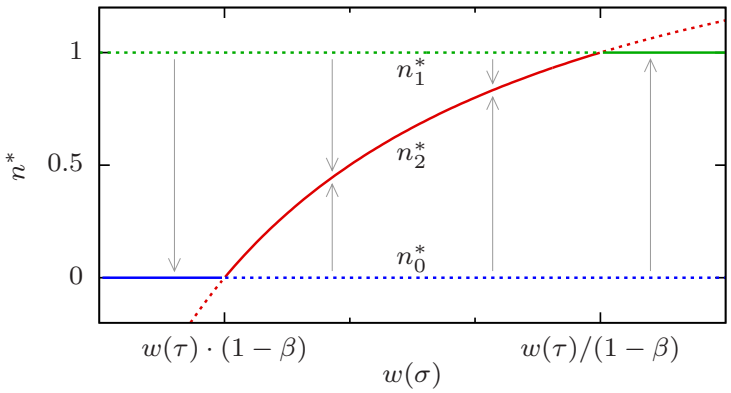



**Figure 5.8.** Fixed point structure of equation (5.9). Solid lines correspond to stable fixed points, dotted lines to unstable ones. Arrows point in the direction of the attractor.

strength $\beta$. A priori it is therefore not clear whether this kind of interactions between individuals leads to an advantage.

Like in case of recombination, the advantage or disadvantage depends on the underlying fitness landscape. As shown in figure 5.9(a) and (c), increasing $\beta$ leads to a disadvantage on the HoC landscape while it is advantageous on the RMF landscape. Though this is qualitatively the same behavior as it was observed for recombining populations, the explanation is very different. In both landscape types, the population adapts faster if interactions become stronger, which can be seen from increased native fitness $\langle w \rangle$. However, since adaptation happens generally very slowly on the HoC landscape, the advantage due to faster adaptation can not compensate for the cost inherent in strong competition. On the RMF landscape, the speed of adaptation is generally much faster if the population is not trapped on local optima, which is largely prevented by frequency-dependent selection with large $\beta$. The further the population gets away from the reference sequence, the larger the fitness becomes on average and hence the ability of strongly competing populations to travel through the landscape faster enhances its advantage by a larger amount than on the HoC landscape. Adaptation is only curbed by the fact that the number of mutations leading uphill decreases with increasing distance traveled. However, this is an effect that vanishes presumably for very large $L$.

The behavior on the different landscape types is also quite different with respect to the diversity $D$. It is not surprising that strong interactions increase the diversity in all cases as shown in figure 5.9(b) and (d). But the diversity converges quickly to a constant value on the HoC landscape, while it seems to grow indefinitely on the RMF landscape. There are at least two explanations for this phenomenon. Either the population moves through the fitness landscape as a bulk with increasing size or the population decomposes into sub-populations that travel more or less independently through the landscape. The answer will be given in the next section where the $L \to \infty$ limit is studied. Due to the

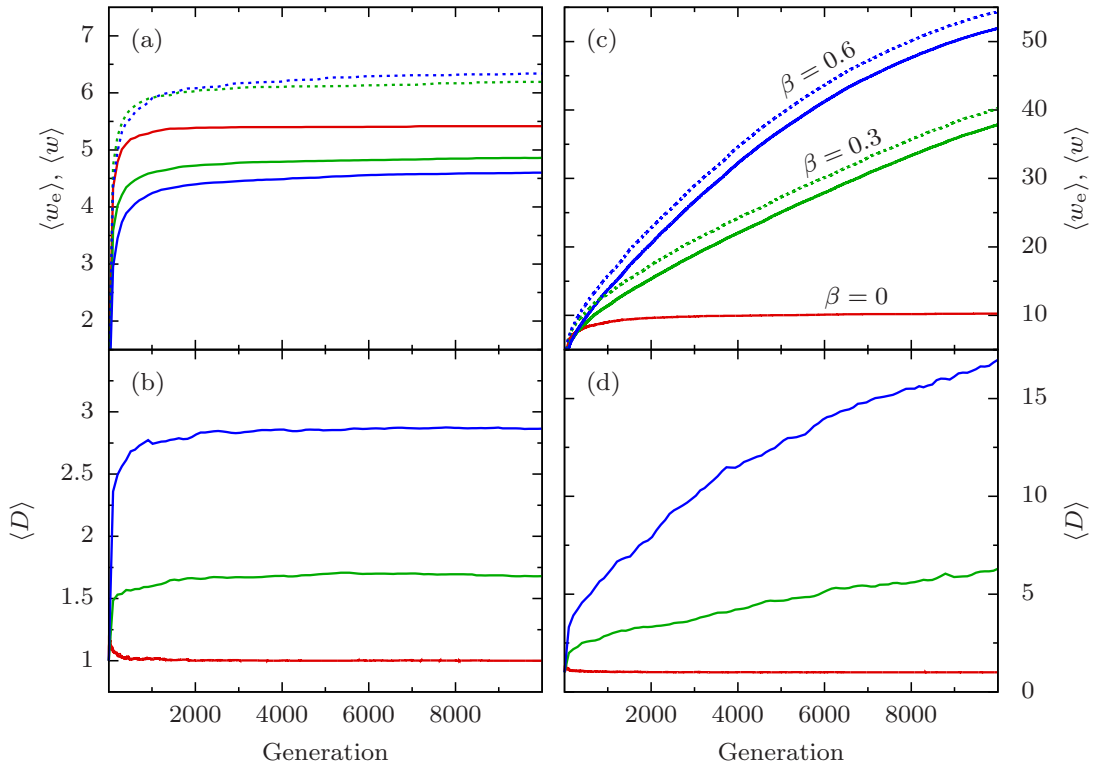



**Figure 5.9.** Averaged time series of mean effective fitness $\langle w_e \rangle$ (solid lines), mean native fitness $\langle w \rangle$ (dotted lines), and diversity $D$. Different colors correspond to different interaction strengths $\beta$ as defined in panel (c). Fitness is assigned according to the RMF model with standard exponentially distributed random part and slope $s = 0$ in (a) and (b) and $s = 1$ in (c) and (d). Other parameters are $L = 64$, $N = 10^5$, and $N\mu = 1$.

finite sequence length $L$ here, the growth of diversity on the RMF landscape will stop eventually when the global optimum is reached.

### 5.3.3. Infinite Genome Limit

Another variant of the Wright-Fisher dynamics described above is supposed to resemble its behavior for the limit $L \to \infty$. In this limit, each mutation produces a new genotype since the probability that an already existing genotype is recreated vanishes and hence the hypercube structure is lost. In order to see how mutations are handled, suppose a monomorphic population that carries the genotype $\sigma$. In the previous version of the model, the mutation step creates $L$ new genotypes with $N\mu/L$ individuals each, which is a small number for large $L$. If the selection term is neglected, random sampling will keep (at least) one individual on each genotype with probability $1 - \exp(-N\mu/L) \approx N\mu/L$. Hence the number of new genotypes is the sum of $L$ Bernoulli random variables with success probability $N\mu/L$, which converges to a Poisson distribution for $L \to \infty$. Therefore, the mutation step in that limit will be replaced by the mere creation of

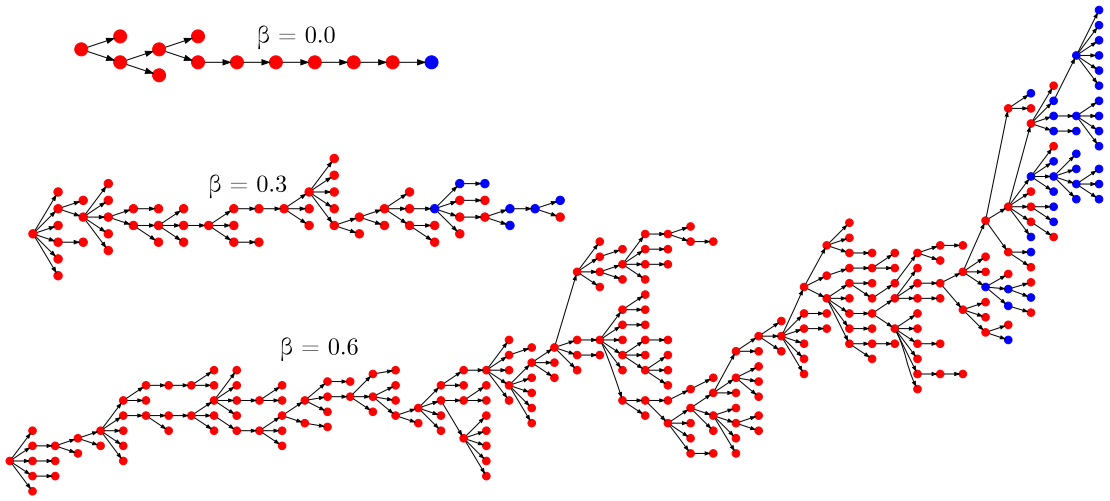



**Figure 5.10.** Phylogenetic trees obtained from the Wright-Fisher model in the infinite genome limit at generation 5000. Each blue dot represents a genotype that was carried by an individual which was alive in the final generation. Red dots represent genotypes with individuals that have produced offspring but became extinct before the end of the simulation. Arrows connect ancestors to descendants. Note that the horizontal alignment of the dots is solely determined by the genotypic relationship and not by the time when the corresponding individuals existed. Fitness is assigned according to the RMF model with $s = 1$ and standard exponentially distributed random part. The population parameters are $N = 10^5$ and $N\mu = 2$.

new genotypes that are carried by single individuals. The number of these genotypes is drawn from said Poisson distribution with parameter $N\mu$. Other steps of the dynamics remain the same. The fitness values are assigned analogously to the RMF model, i.e., by a random part plus a term proportional to the number of mutations, with $s$ being the constant of proportionality.

In this scenario, the adaption process is not influenced by a changing geometry of the underlying landscape. Furthermore, each genotype has an unique ancestral genotype such that the genealogical relationships can always be represented by trees. Examples of such trees are shown in figure 5.10. As one can see, their structure is considerably influenced by the interaction strength $\beta$. For $\beta = 0$, i.e., without interactions, the resulting trees are almost linear with very few and short ramifications. If $\beta$ is increased, the trees grow in both height and width. Since all trees correspond to the 5000th Wright-Fisher generation, the difference in height is simply due to a faster adaptation process associated with stronger interactions. More importantly, the number of branches increases as well and one finds for strong interactions of $\beta = 0.6$ many coexisting individuals that have a rather large genetic distance. This suggests that the populations indeed splits into multiple sub-populations that adapt independently, at least on smaller timescales. On longer timescales, however, one observes that only one of the parallel branches survives and hence the most recent common ancestor of coexisting individuals is genetically rather close compared to the total height of the tree. Furthermore, a

consequence of the RMF model is that the mean fitness of the population will grow indefinitely and hence living individuals are usually close to the leaves of the tree.

# 6. Discussion

## 6.1. Summary

In this thesis, mathematical models for high dimensional fitness landscapes and evolutionary processes thereon were studied. Three landscape models were covered, the House-of-Cards (HoC), Rough-Mount-Fuji (RMF), and NK model, which differ in most quantities under consideration. The latter two have a parameter (denoted by $s$ and $K$, respectively, in this thesis) that allows for the interpolation between a smooth landscape and the maximally rugged HoC landscape. Apart from that, NK models have many additional degrees of freedom due to the choice of the epistatic interaction pattern that affects the overall landscape topography in a rather subtle way. In comparison, the HoC landscape does not have a particular high complexity, but it largely facilitates pure analytical studies.

Chapter 2 and 3 addressed the ruggedness of fitness landscape with two main approaches: Local maxima and accessible paths. This was done independent of any notion of evolutionary dynamics, i.e., the results are largely applicable to general value landscape as well and not only to fitness landscapes in the biological context. In chapter 4, a rather simple evolutionary model, the adaptive walk, was introduced in order to study how the landscape influences the dynamics. Finally, the more sophisticated Wright-Fisher dynamics was studied in chapter 5. It revealed for two generic examples, recombination and competition, that properties of the underlying fitness landscape are crucial for certain question in evolutionary biology.

### 6.1.1. Local Maxima and Rank

A first direct proxy for the landscape ruggedness that was discussed is the number $N_{\max}$ of maxima. Closely related is the question which fitness $h$ a typical maximum has. On the HoC landscape, these questions are almost trivial to answer as the statistics of a local maximum's fitness are the same as for the largest of $L+1$ i.i.d. random variables. Furthermore, one can show that local maxima tend slightly to form clusters. The situation on the RMF landscape is already more complicated. In the general case, $N_{\max}$ can be written as a one-dimensional integral (equation (2.31)), which can be expressed in simple closed form for a few distributions of the random part. If the distribution's tail is too heavy, the fitness is mostly influenced by fluctuations rather than the overall fitness gradient. As a consequence, the leading order behavior of $N_{\max}$ is the same as for the HoC landscape, independent of the value of the gradient's slope $s$. Another notable feature of this landscape type is that the probability to find local maxima increases with

distance to the reference sequence $\tilde{\sigma}$. Therefore, they are not evenly distributed over the landscape but usually clustered in the proximity of the global maximum.

A general expression for $N_{\max}$ in case of the NK model can be written as an $L$-dimensional integral (equation (2.38)) that involves the $k$-fold convolution of a distribution for arbitrary $k$. The same applies to the fitness $h$ of a typical maximum. Like for the RMF model, the expressions can be evaluated for a few special cases. An important example is the block model, where the interaction scheme is chosen in a way that genotypes are divided into independent blocks. Therefore, both $N_{\max}$ and $\mathbb{E}[h]$ can be computed simply from already known results of the HoC model. One can also obtain these quantities analytically for adjacent interactions, certain distributions, and small values of $K$. Random interactions still rely solely on numerical analysis. They reveal that $N_{\max}$ is smallest for random interactions and largest for block interactions in all cases considered here. Adjacent interactions are generally somewhere in between. For the mean fitness $\mathbb{E}[h]$ of maxima, this order is reversed. A helpful measure to quantify arbitrary interaction schemes is the rank $R$, which is strongly correlated to most NK landscape properties including $N_{\max}$ and $\mathbb{E}[h]$. Generally, the landscape becomes less rugged if the rank is increased while keeping the landscape parameters $L$ and $K$ constant.

### 6.1.2. Accessible Paths

The study of accessible paths to the global maximum $\hat{\mathbf{1}}$ from its antipodal sequence $\hat{\mathbf{0}}$ is more complex than local maxima as they have to cross the whole landscape by definition. In contrast to local maxima, which can be studied on a rather small area of the landscape, one has to take the global structure of the landscape into account in order to study accessibility. For this reason, the most detailed results were obtained for the HoC model.

On this landscape, the probability for the accessibility of a given shortest path decays as $1/L!$ and is therefore generally very small. This is partially compensated by the fact that there are also $L!$ shortest paths available such that the mean number of accessible paths is given by $\mathbb{E}[X] = 1$. Due to the hypercube structure, the fitness of the initial genotype $\hat{\mathbf{0}}$ has crucial influence on the probability $\mathbb{P}[X > 0]$ that there is at least one accessible path since it is the only genotype that is contained in every possible path (apart from the global optimum $\hat{\mathbf{1}}$). This results in a phase transition with regard to the initial fitness where $\lim_{L\to\infty} \mathbb{P}[X > 0]$ jumps from 0 to 1. On the directed hypercube, where only shortest paths are allowed, the transitions happens around $w(\hat{\mathbf{0}}) = \log(L)/L$. In case of the undirected hypercube, where arbitrary paths are allowed, the transition is at a constant value of $w(\hat{\mathbf{0}}) \approx 0.1186$.

Furthermore, most results suggest that the directed hypercube has generally a critical topography in the sense that small changes have a notable effect on the percolation probability. When the initial fitness is randomly chosen, the probability $\mathbb{P}[X > 0]$ tends to 0. For the undirected hypercube, the limiting probability is already given by a positive constant. If the analogous problem is studied on a tree, there is even a phase transition with regard to the branching number $n$ where $\mathbb{P}[X > 0]$ jumps again from 0 to 1. In order to mimic the hypercube, the branching number should be scaled such that the

total number of paths is asymptotically given by $L!$. This is the case for $n = e\,L$, which is also the critical value where the said transition happens.

Another way of increasing the accessibility is to introduce a global fitness gradient as it is the case in the RMF model. On the regular tree with fixed branching number $n$, $\mathbb{P}[X > 0]$ shows a continuous phase transition with respect to $s$. On the hypercube, any positive slope $s$ of the gradient is sufficient to raise $\mathbb{P}[X > 0]$ to 1, even if $s$ converges to zero sufficiently slowly as $L \to \infty$. Conversely, the probability $\mathbb{P}[X > 0]$ goes to zero in the NK model for any classic interaction pattern and fixed $K > 1$. This is somewhat counter-intuitive as the ruggedness in terms of local maxima is still lower than for, e.g., RMF landscapes with small $s$ for which $\mathbb{P}[X > 0] \to 1$. However, numerically obtained values show that $\mathbb{P}[X > 0]$ is correlated with the interaction rank $R$. As $N_{\max}$ is negatively correlated with $R$, both proxies for the ruggedness are consistent within the class of NK landscapes. Note also that detailed results on $X$ for blockwise interactions can be obtained from results for the HoC model.

### 6.1.3. Adaptive Walks

During the study of adaptive walks, the quantities of interest were walk length $\ell$ and the final fitness $h$. Once more, the distribution of both quantities can be obtained analytically in case of the HoC landscape and adaptive walks of greedy, random and reluctant type. Natural adaptive walks do not quite fit in the scheme as their behavior depends on the fitness distribution, especially on its tail. However, the tail can be represented by the generalized Pareto distribution for which the leading order of the mean walk length is known analytically.

The intuition that walks last longer on less rugged landscape is mostly confirmed by the corresponding study on RMF and NK landscape. As already mentioned, the behavior of $N_{\max}$ changes when the fluctuations of the random part of the RMF model are of the same order as the slope $s$. This behavior is also inherited by random adaptive walks since they behave like on a HoC landscape if the fluctuations are larger and like on an additive landscape if the fluctuations are smaller. If the distribution of the random part is exponential, there is a phase transition with respect to $s$ where the walk length $\ell$ switches from logarithmic to linear behavior.

NK landscapes and their tunable interaction patterns also reveal effects of more subtle changes of the landscape. It was shown that adaptive walks through a landscape with blockwise interactions can be interpreted as a set of independent walks through HoC landscapes. As a consequence, the walk statistics can be obtained once more from HoC results. In the general case, the length of greedy walks is roughly proportional to the mean distance between two maxima, but random and reluctant walks are also affected by landscape properties that cannot be ascribed to the maxima density only. Especially the combination of random interactions and reluctant walks shows peculiar behavior. Their walk length depends non-monotonically on the ruggedness parameter $K$ and can be several times larger than the genome size $L$, i.e., each locus is flipped multiple times before the walk terminates. More surprisingly, reluctant walks reach a higher

average fitness than random and greedy walks for intermediate values of $K$ and random interactions, even though it is the other way round on most other landscape types.

### 6.1.4. Wright-Fisher Dynamics

The consequences of landscape ruggedness on the Wright-Fisher dynamics are very complex. Generally, adaptation happens faster the smoother the underlying landscape is, at least in the scenarios considered here. However, the details of the response to different landscapes depends on the exact dynamics under consideration.

For dynamics that includes recombination, it was shown that it is superior to asexual dynamics under certain circumstances. The problem that recombining populations have is that they get trapped easily on local maxima where they cannot escape from. However, before this happens, they can adapt faster than asexual ones. This results in a non-monotonic temporal pattern of the fitness advantage that recombining populations have. On very short and long timescales, recombination is almost always disadvantageous, but there might be an advantageous time frame at intermediate times. Landscape ruggedness now comes into play due to the fact that it determines how long it takes to get trapped on a local maximum and how large the advantage is that recombining populations build up. The less rugged a landscape is, the longer is this time window and the bigger is the advantage. There are also several mechanism that prolong and increase the advantage. On a time-dependent landscape, where local maxima exist only temporarily, the advantage can be prolonged even indefinitely in accordance with the Red-Queen hypothesis.

The situation is similar for the dynamics with competition, which was modeled via frequency-dependent selection. In the scenarios considered here, competitions lead to faster adapting populations, while they decrease the effective fitness. Whether the latter can be compensated by fast adaptation depends again on the underlying landscape. Probably the more interesting aspect of this model is its production of genetic diversity. It was shown that the dynamics allows for the split of populations into independently evolving sub-populations. This leads to complex phylogenetic trees for sufficiently large genome sizes $L$.

## 6.2. Open Questions and Outlook

Many questions related to the HoC model are already answered. With regard to local maxima, the asymptotic distribution of the number $N_{\max}$ of local maxima is known [64, 106] as well as the probability that two sequences at a certain distance are both maxima. The same applies to the distribution of the number $X$ of shortest accessible path if scaled properly. A related question that arises frequently is what happens when the requirement for accessibility is loosened [107, 108]. For instance, one can allow for a certain number of steps along the path that lower the fitness, that the accumulated fitness loss caused by these steps is below a certain threshold $\epsilon$, or a combination of both. Note that a version where the fitness is allowed to be lowered by an amount smaller than $\epsilon$ in every step is equivalent to the RMF model with slope $s = \epsilon$

on the directed hypercube, but it makes a difference in the undirected case. Concerning the RMF model on the directed hypercube, it was shown that it is always accessible if $s$ is constant or goes to 0 with $L$ more slowly than $1/L$, but apart from that it is not known how $s$ can be scaled with $L$ such that a transition from high to low accessibility becomes visible.

The concept of accessibility percolation can also be applied to graphs different from hypercubes and trees. In fact, the HoC and RMF model together with the corresponding notion of accessibility can be applied to any graph. A biologically relevant example are Hamming graphs, i.e., the analog to hypercubes for an alphabet size larger than $\{0, 1\}$. With this, one can also model DNA sequences consisting of 4 letters or protein landscapes consisting of 20 letters. However, the problem has to be rephrased slightly since the antipodal sequence of the global maximum (or any other sequence) is not defined uniquely in that case. Another example outside biology is the graph where vertices correspond to permutations which can be used, for instance, to treat the traveling salesman problem [10, 11]. In the latter case, it would make sense to use the landscape associated with that problem [11] rather than the HoC or RMF model.

Regarding the NK model and its interaction schemes, there are many open questions related to the number of maxima, accessibility and adaptive walks. As mentioned above, $N_{\max}$ and $\mathbb{E}[h]$ are only known for a few special cases including blockwise and adjacent interactions, but the latter only for small values of $K = 2$ or $K = 3$. The behavior of adjacent or random interactions with arbitrary values of $K$ is not known analytically, nor how the distribution of fitness contributions affects landscape properties. This could be examined with help of equation (2.38) in future work. Additionally, one can presumably use the ansatz that led to this equation in order to derive an equation for the probability that two sequences at a certain distance are both maxima. More importantly, there is no analytical approach that explains how the interaction patterns affect the landscape. As shown, most properties under consideration are correlated with the rank $R$ that can be computed analytically, but it has only a describing function rather than being able to explain the behavior. Moreover, little is known about the number $X$ of accessible paths through the NK landscape. It can be shown that $\mathbb{P}[X > 0] \to 0$ as $L \to \infty$ for fixed $K$, but it is not known how the percolation probability behaves when $K$ is scaled with $L$. The same applies to the distribution of $X$, which can only be computed for blockwise interactions. This interaction type is also the only one for which the length and height of adaptive walks can be computed exactly. The walk behavior is influenced by rather complicated features of the landscape that go beyond the sheer number of maxima. In particular, an explanation is still lacking for the phenomenon that random and reluctant walks can reach higher fitness than greedy walks under certain circumstances.

For the Wright-Fisher dynamics, there is such a vast number of possible scenarios that not all of them can be listed here, especially with regard to recombination. One example is the fact that the models used here are only suitable for haploid populations. The consideration of diploid genotypes might have interesting consequences, e.g., equilibria of a population in the deterministic limit do not coincide strictly with local maxima of the landscape anymore [109]. Related to the study of adaptive walks in this thesis, a careful examination of the Wright-Fisher dynamics on NK landscapes could assess how

important the interaction pattern is for more complex dynamics. Moreover, the version including frequency-dependent selection can be studied in much more detail to obtain, e.g., statistics about the phylogenetic trees produced by this model. In order to make it more realistic, one could also define a reasonable mapping from genotypes to phenotypes such that that frequency-dependence acts only on the latter. As mentioned earlier, the diversity creating part of competition is also a good way to improve the advantage of recombination. Therefore, it is convenient to study the model where both recombination and competition are combined.

# A. Appendix

## A.1. Notation and Definitions

### A.1.1. Symbols and Functions

| | |
|---|---|
| $\hat{\mathbf{0}}$, $\hat{\mathbf{1}}$ | The sequences consisting only of zeroes and ones, respectively, i.e., $\hat{\mathbf{0}} = (0, \ldots, 0)$ and $\hat{\mathbf{1}} = (1, \ldots, 1)$. |
| $\mathbb{E}[V]$ | Expected value of the random variable $V$. |
| $\langle V \rangle$ | The average of a quantity $V$. Used in a broader sense than $\mathbb{E}[V]$. The exact meaning is explained at the corresponding passage in the text. |
| $f_1 \sim f_2$ | The functions or sequences $f_1$ and $f_2$ are asymptotically equivalent, i.e., $\lim_{x\to\infty} f_1(x)/f_2(x) = 1$. |
| $f_1 = \mathcal{O}(f_2)$ | The function or sequence $f_2$ grows at most as fast as $f_2$, i.e., $\limsup_{x\to\infty} f_1(x)/f_2(x) < \infty$. |
| $\Gamma(x)$ | Gamma function defined by $\Gamma(x) = \int_0^\infty y^{x-1} e^{-y}\,\mathrm{d}y$. |
| $I_{[E]}$ | Indicator random variable of the event $E$, i.e., $I_{[E]} = 1$ if $E$ happens and $I_{[E]} = 0$ otherwise. |
| $I_S(x)$ | Indicator function of the set $S$, i.e., $I_S(x) = 1$ if $x \in S$ and $I_S(x) = 0$ otherwise. |
| $\log(x)$ | The natural logarithm of $x$. |
| $\mathbb{P}[E]$ | Probability of the event $E$. |
| $\theta(x)$ | Heaviside function, i.e., $\theta(x) = 1$ for $x > 0$ and $\theta(x) = 0$ otherwise. |
| $w(\sigma)$ | Fitness of genotype $\sigma$. |

Furthermore, a string of numbers in monospace font (e.g., `12345`) encodes a path through the hypercube as described in section 2.1.

### A.1.2. Variables and Parameters

| | |
|---|---|
| $h$ | Fitness of a local maximum. |
| $K$ | Ruggedness parameter of the NK model. |
| $\ell$ | Number of steps during an adaptive walk. |
| $L$ | Number of loci / Dimension of the hypercube. |
| $n$ | Branching number of regular trees. |
| $N$ | Population size. |
| $N_{\max}$ | Mean number of local maxima. |
| $\mu$ | Mutation rate. |
| $P_{\max}$ | Probability that a certain sequence is a local maxima. |
| $R$ | Rank of epistatic interactions. |
| $s$ | Slope of the fitness gradient of RMF landscapes. |
| $X$ | Number of accessible paths to the global maximum. |

### A.1.3. Abbreviations

| | |
|---|---|
| AW | Adaptive walk |
| CDF | Cumulative distribution function |
| HoC | House of Cards |
| i. i. d. | Independent and identically distributed |
| PDF | Probability density function |
| RMF | Rough Mount Fuji |

### A.1.4. Probability Distributions

Let $Z$ denote the random variable distributed according to the corresponding distribution. If applicable, $f$ denotes its PDF.

| | |
|---|---|
| Bernoulli distribution | $Z = 1$ with probability $p$ and $Z = 0$ with probability $1 - p$. |
| Binomial distribution | $\mathbb{P}[Z = k] = \binom{M}{k} p^k (1-p)^{M-k}$ for $k \in \{0, 1, \ldots, M\}$, where $M$ is a positive integer. |

Exponential distribution $f(x) = 1/\lambda \exp(-x/\lambda)\,\theta(x)$, where $\lambda$ is the mean value. The standard exponential distribution corresponds to $\lambda = 1$.

Gamma distribution $f(x) = \dfrac{x^{p-1}e^{-\frac{x}{b}}}{b^p\Gamma(p)}\,\theta(x)$, where $p > 0$ is the shape parameter and $b > 0$ is the scale parameter.

Generalized Pareto distr. $f(x) = \begin{cases} (\kappa x + 1)^{-(\kappa+1)/\kappa}\,\theta(x) & \text{for } \kappa > 0, \\ (\kappa x + 1)^{-(\kappa+1)/\kappa}\,I_{[0,-1/\kappa]}(x) & \text{for } \kappa < 0, \\ e^{-x}\,\theta(x) & \text{for } \kappa = 0. \end{cases}$

Gumbel distribution $f(x) = \exp\bigl[-(x + e^{-x})\bigr]$.

Normal distribution $f(x) = \dfrac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left[-\dfrac{(x-\mu)^2}{2\sigma^2}\right]$. The standard normal distribution corresponds to $\mu = 0$ and $\sigma = 1$.

Poisson distribution $\mathbb{P}[Z = k] = \dfrac{\lambda^k}{k!}\exp(-\lambda)$ for $k \in \mathbb{N}_0$, where $\lambda$ is the mean value.

Uniform distribution $f(x) = \dfrac{1}{b-a}\,I_{[a,b]}(x)$.

## A.2. Algorithms

### A.2.1. Creation of Interaction Patterns with Arbitrary Rank

In order to create interaction schemes with arbitrary rank, the following algorithm is used:

1. Set the interaction sets $V_i$ to blockwise interactions.
2. Define a target rank $r_{\text{aim}}$.
3. Draw a random number $i \in \{1, \ldots, L\}$, choose a random element $j \in V_i$ with $j \neq i$ and replace it with a randomly chosen element $j' \notin V_i$.
4. If the current rank is closer to the target rank, accept the change of step 3, otherwise undo it.
5. If the change is rejected 1000 times in a row, the algorithm will be aborted. Start again at step 1.
6. If the relative distance to the target rank is smaller than 1%, accept it for further study. Go to step 2.

It is convenient to increase the target rank $r_{\text{aim}}$ after each successfully created interaction scheme such that the pattern changes smoothly from a block-like structure to a disordered one until the algorithm gets aborted in step 5. The results are sufficient in order to interpolate between low an high rank. Note, however, that the algorithm will usually reach ranks of at most $R_{\text{rnd}}$ rather than $R_{\max}$.

### A.2.2. Numerical Computation of $P_{\max}$ and $\mathbb{E}[h]$ for the NK Model

The probability $P_{\max}$ that a randomly chosen genotype $\sigma$ is a maximum is given by equation (2.38). It is basically the $L$-dimensional integral of the integrand

$$g(\mathbf{x}) = \prod_{i=1}^{L} \Big[ f(x_i)\, \tilde{F}_{|U_i|}\Big( \sum\nolimits_{j \in U_i} x_j \Big) \Big] \,. \tag{A.1}$$

The computation of the integral is based on Monte-Carlo integration or, more precisely, on the importance sampling algorithm [110]. In general, an estimator for the integral is given by

$$I = \int g(\mathbf{x})\, \mathrm{d}^L\mathbf{x} \approx \sum_{i=1}^{n} \frac{g(\mathbf{x}_i)}{p(\mathbf{x}_i)} = I_n \,, \tag{A.2}$$

where $p(\mathbf{x})$ is a probability density and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are points randomly drawn from the corresponding distribution. The error is given by

$$\Delta I_n = \sqrt{\frac{1}{n\,(n-1)} \sum_{i=1}^{n} \left( \frac{g(\mathbf{x}_i)}{p(\mathbf{x}_i)} - I_n \right)^2} \,. \tag{A.3}$$

Of course, $p$ should be chosen such that $\Delta I_n$ is as small as possible. This, in turn, is achieved when $p$ has a shape similar to $g$. For the integrand defined by equation (A.1), a good choice is given by

$$p(\mathbf{x}) = \prod_{i=1}^{L} f(x_i - x_i^{\max}) \,, \tag{A.4}$$

where $\mathbf{x}^{\max}$ is the maximum of $g$. Then most sample points are in a region where the integrand has most of its weight. Since $g$ is usually a single peaked function, the maximum can be found easily by a simple hill-climbing algorithm. The number $n$ of samples is adjusted during the process such that the relative error $\Delta I_n / I_n$ is below a certain threshold. In all figure that involve $P_{\max}$ shown in this thesis, the error is much smaller than the symbol size.

This works analogously for $\mathbb{E}[h]$ by taking $g$ to be the integrand of equation (2.39). Note that the error is noticeably larger in that case, because the weight function $p$ given by equation (A.4) differs more from $g$.

# A.3. Proof of Equation (3.20)

## A.3.1. Improving the Lower Bound on $\mathbb{P}[X_L > 0]$

Let $n = \alpha L$ with $\alpha > e^{-1}$ in the following. According to equation (3.15) and the fact that $\mathbb{E}[X_L]$ serves as an upper bound for $\mathbb{P}[X_L > 0]$, it follows immediately that $\mathbb{P}[X_L > 0] \to 0$ for $\alpha < e^{-1}$. In order to show that $\mathbb{P}[X_L > 0] \to 1$ for $\alpha > e^{-1}$ one has to improve the lower bound. The strategy in [48] was to introduce paths that are accessible and fulfill additionally the condition that the $i$-th vertex $\sigma_i$ after the root has fitness of at least $w(\sigma_i) > \epsilon + (1-\epsilon)(i-1)/L$. In the following, the set of those paths will be denoted by $D_\epsilon$ and their number by $X_{L,\epsilon} = |D_\epsilon|$. Obviously $X_{L,\epsilon} < X_L$ since they have to fulfill an additional condition compared to ordinary accessible paths. Therefore, according to (3.10)

$$\mathbb{P}[X_L > 0] \geq \mathbb{P}[X_{L,\epsilon} > 0] \geq \frac{\mathbb{E}[X_{L,\epsilon}]^2}{\mathbb{E}\left[X_{L,\epsilon}^2\right]} \tag{A.5}$$

which will turn out to be a sharper lower limit for $\mathbb{P}[X_L > 0]$ than $\mathbb{E}[X_L]^2/\mathbb{E}\left[X_L^2\right]$.

To calculate the first moment, let $\Sigma$ be an arbitrary path from the root to a leaf. For $\epsilon = 0$, the probability that $\Sigma$ is contained in $D_0$ can be written as

$$\mathbb{P}[\Sigma \in D_0] = \int_{\frac{L-1}{L}}^{1} \mathrm{d}x_L \int_{\frac{L-2}{L}}^{x_{L-1}} \mathrm{d}x_{L-1} \ldots \int_{\frac{1}{L}}^{x_3} \mathrm{d}x_2 \int_{0}^{x_2} \mathrm{d}x_1 \,. \tag{A.6}$$

This chain of integrals can be formulated equivalently as a recurrence relation

$$I_j(x) = \int_{\frac{j-1}{L}}^{x} I_{j-1}(y)\,\mathrm{d}y \tag{A.7}$$

with $I_0(x) = 1$ and $\mathbb{P}[\Sigma \in D_0] = I_L(1)$. The solution of (A.7) reads

$$I_j(x) = \frac{(Lx - n + 1) \cdot (Lx + 1)^{n-1}}{j! \cdot L^n} \tag{A.8}$$

and hence

$$\mathbb{P}[\Sigma \in D_0] = I_L(1) = \frac{(L+1)^{L-1}}{L! \cdot L^L} \geq \frac{1}{L \cdot L!} \,. \tag{A.9}$$

The corresponding result for $\epsilon > 0$ is straightforward to obtain, since uniform random variables on $[0, 1]$, conditioned on being larger than $\epsilon$, are simply uniformly distributed on $[\epsilon, 1]$. This leads to

$$\begin{aligned} \mathbb{P}[\Sigma \in D_\epsilon] &= \mathbb{P}[\Sigma \in D_\epsilon \mid \forall\, i : w(\sigma_i) > \epsilon] \cdot \mathbb{P}[\forall\, i : w(\sigma_i) > \epsilon] \\ &= \mathbb{P}[\Sigma \in D_0] \cdot \mathbb{P}[\forall\, i : w(\sigma_i) > \epsilon] = \frac{(1-\epsilon)^L}{L \cdot L!} \end{aligned} \tag{A.10}$$

and with Stirling's formula to

$$\mathbb{E}[X_{L,\epsilon}] = n^L \, \mathbb{P}[\Sigma \in D_\epsilon] = \frac{[n(1-\epsilon)]^L}{L \cdot L!} \geq \frac{z^L}{e \, L^{3/2}} \,, \tag{A.11}$$

where $z = \alpha \, (1-\epsilon) \, e$.

To obtain the second moment, the probability $p_{k,\epsilon}$ that two paths sharing $k+1$ vertices are both in $D_\epsilon$ is needed. Like before, consider the case $\epsilon = 0$ first which can be generalized easily to $\epsilon > 0$ afterwards. The calculation works similar to that of equation (3.7), but now the fitness $x$ of the last common vertex has to be at least $(k-1)/L$. Omitting the additional constraints on the other vertices yields for $k > 1$

$$p_{k,0} \leq \int_{\frac{k-1}{L}}^{1} \frac{x^{k-1}}{(k-1)!} \left( \frac{(1-x)^{L-k}}{(L-k)!} \right)^2 \mathrm{d}x \,. \tag{A.12}$$

Using Stirling's formula and the fact that $x^{k-1} \, (1-x)^{2L-2k}$ has its maximal value at $(k-1)/(2L-k-1) < (k-1)/L$ leads to

$$\begin{aligned} p_{k,0} &\leq \frac{\left(\frac{k-1}{L}\right)^{k-1}}{(k-1)!} \left( \frac{\left[1 - \left(\frac{k-1}{L}\right)\right]^{L-k}}{(L-k)!} \right)^2 \\ &\leq \left(\frac{e}{L}\right)^{2L-k-1} \cdot \frac{[(1+L-k)/(L-k)]^{2L-2k}}{(2\pi)^{\frac{3}{2}} \sqrt{k-1} \, (L-k)} \\ &\leq \frac{1}{(L-k)} \left(\frac{e}{L}\right)^{2L-k-1} \end{aligned} \tag{A.13}$$

for $k > 1$ and

$$p_{1,0} \leq \frac{1}{(2L-1) \cdot (L-1)!^2} \leq \left(\frac{e}{L}\right)^{2L} \quad \text{and} \quad p_{0,0} = \left(\frac{1}{L \cdot L!}\right)^2 \tag{A.14}$$

for $k = 1$ and $k = 0$, respectively. Like before, the case $\epsilon > 0$ is obtained by multiplying the result with the probability that all vertices are larger than $\epsilon$, i.e., $p_{k,\epsilon} = (1-\epsilon)^{2L-k} p_{k,0}$. The second moment is then at most

$$\begin{aligned} \mathbb{E}\big[X_{L,\epsilon}^2\big] &\leq \mathbb{E}[X_{L,\epsilon}] + \sum_{k=0}^{L-1} m_k \, p_{k,\epsilon} \\ &\leq \mathbb{E}[X_{L,\epsilon}] + \mathbb{E}[X_{L,\epsilon}]^2 + z^{2L} + \alpha L \sum_{k=2}^{L-1} \frac{z^{2L-k-1}}{L-k} \,. \end{aligned} \tag{A.15}$$

Provided that $z = \alpha \, (1-\epsilon) \, e > 1$, which can always be fulfilled for sufficiently small $\epsilon$ if $\alpha > e^{-1}$, there exist a constant $c > 0$ such that

$$c \cdot \sqrt{z^k} \geq k+1 = \left(1 - \frac{k}{k+1}\right)^{-1} \geq \left(1 - \frac{k}{L}\right)^{-1} = \frac{L}{L-k}$$

for all $1 < k < L$ and hence

$$\alpha L \sum_{k=2}^{L-1} \frac{z^{2L-k-1}}{L-k} \leq \alpha z^{2L} \sum_{k=2}^{L-1} \frac{L}{z^k(L-k)} \leq \alpha c z^{2L} \sum_{k=2}^{\infty} \frac{1}{\sqrt{z^k}} . \tag{A.16}$$

Thus the upper bound for $\mathbb{E}\Big[X_{L,\epsilon}^2\Big]$ can then be written as

$$\mathbb{E}\big[X_{L,\epsilon}^2\big] \leq \mathbb{E}[X_{L,\epsilon}] + \mathbb{E}[X_{L,\epsilon}]^2 + c' \cdot z^{2L} \tag{A.17}$$

and combining this with equations (3.10) and (A.11) leads to

$$\mathbb{P}[X_{L,\epsilon} > 0] \geq \frac{1}{\mathbb{E}[X_{L,\epsilon}]^{-1} + 1 + c' \cdot e\, L^3} \geq \frac{\tilde{c}}{L^3} \tag{A.18}$$

for some positive constants $c'$ and $\tilde{c}$.

### A.3.2. Raising the Lower Bound to One

Now that it is shown that $\mathbb{P}[X_{L,\epsilon} > 0]$ decays at most as $1/L^3$, the last step is to improve this bound to something that converges to one. The idea is to compensate this decay by showing that there are so many paths up to the fourth level of the tree with fitness smaller than $\epsilon$ that at least one of them can be extended to the $L$-th level for $L \to \infty$. Let $\tilde{X}_{j,\epsilon}$ denote the number of paths that go from the root to the $j$-th level and fulfill that the $i$-th vertex $\sigma$ after the root has fitness $w(\sigma) \in [(i-1)/(4\epsilon), i/(4\epsilon)]$. To begin with, $\tilde{X}_{1,\epsilon}$ can be interpreted as the sum of $n$ independent Bernoulli distributed random variables, each of them having probability $\epsilon/4$ to be equal to 1 and hence $\mathbb{E}\Big[\tilde{X}_{1,\epsilon}\Big] = n\,\epsilon/4$. With Chernoff's bound (3.41) it follows that

$$\mathbb{P}\Big[\tilde{X}_{1,\epsilon} < \frac{n\,\epsilon}{8}\Big] \leq \exp\Big(-\frac{n\epsilon}{32}\Big) .$$

Now given that there are at least $(n\epsilon/8)^j$ paths to the $j$-th level, $\tilde{X}_{j+1,\epsilon}$ is again the sum of at least $n \cdot (n\epsilon/8)^j$ Bernoulli random variables and hence

$$\mathbb{E}\Big[\tilde{X}_{j+1,\epsilon} \mid \tilde{X}_{j,\epsilon} > (n\epsilon/8)^j\Big] \geq n \cdot \Big(\frac{n\epsilon}{8}\Big)^j \cdot \frac{\epsilon}{4} = 2\Big(\frac{n\,\epsilon}{8}\Big)^{j+1}$$

Applying the Chernoff bound again leads to

$$\mathbb{P}\Big[\tilde{X}_{j+1,\epsilon} < (n\epsilon/8)^{j+1} \mid \tilde{X}_{j,\epsilon} > (n\epsilon/8)^j\Big] \leq \exp\Big[\frac{1}{4}\Big(\frac{n\,\epsilon}{8}\Big)^{j+1}\Big] . \tag{A.19}$$

For the sake of a shorter notation, let $A_j$ be the event that $\tilde{X}_{j,\epsilon} < (n\epsilon/8)^j$. Then,

$$\begin{aligned}
\mathbb{P}[A_4] &= \mathbb{P}[A_4 \mid A_3]\,\mathbb{P}[A_3] + \mathbb{P}[A_4 \mid \neg A_3]\,\mathbb{P}[\neg A_3] \leq \mathbb{P}[A_4 \mid \neg A_3] + \mathbb{P}[A_3] \\
&\leq \ldots \leq \mathbb{P}[A_4 \mid \neg A_3] + \mathbb{P}[A_3 \mid \neg A_2] + \mathbb{P}[A_2 \mid \neg A_1] + \mathbb{P}[A_1] \\
&\leq 4 \exp\Big(-\frac{n \cdot \epsilon^4}{16384}\Big)
\end{aligned}$$

and accordingly

$$\begin{aligned}\mathbb{P}[X_{L,\epsilon}=0] &= \mathbb{P}[X_{L,\epsilon}=0 \wedge A_4] + \mathbb{P}[X_{L,\epsilon}=0 \wedge \neg A_4] \\ &\leq \mathbb{P}[A_4] + \mathbb{P}\Big[X_{L,\epsilon}=0 \mid \tilde{X}_{4,\epsilon} > (n\epsilon)^4/4096\Big]\,. \end{aligned} \tag{A.20}$$

Given that there are $(n\epsilon)^4/4096$ vertices at level four with fitness smaller than $\epsilon$, each of them may serve as the root of a sub-tree of height $L-4$ from which a path in $D_\epsilon$ originates. Using inequality (A.18) finally yields

$$\begin{aligned}\mathbb{P}[X_L=0] \leq \mathbb{P}[X_{L,\epsilon}=0] &\leq \mathbb{P}[A_4] + \mathbb{P}[X_{L-4,\epsilon}=0]^{(n\epsilon)^4/4096} \\ &\leq 4\exp\left(-\frac{n\cdot\epsilon^4}{16384}\right) + \left(1-\frac{\tilde{c}}{(L-4)^3}\right)^{(n\epsilon)^4/4096} \xrightarrow{L\to\infty} 0\,. \end{aligned} \tag{A.21}$$

# Bibliography


[1] C. Darwin. *On the origins of species by means of natural selection* (John Murray, London, 1859).

[2] S. Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress on Genetics*, vol. 1, pp. 355–366 (1932).

[3] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug and J. A. G. M. de Visser. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2013 p. P01005 (2013).

[4] J. A. G. M. de Visser and J. Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, vol. 15 pp. 480–490 (2014).

[5] J. A. G. de Visser, T. F. Cooper and S. F. Elena. The causes of epistasis. *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 278 pp. 3617–3624 (2011).

[6] D. M. Weinreich, R. A. Watson, L. Chao and R. Harrison. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, vol. 59 pp. 1165–1174 (2005).

[7] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich and S. J. Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, vol. 445 pp. 383–386 (2007).

[8] D. L. Stein. *Spin Glasses and Biology* (World Scientific, Singapore, 1992).

[9] E. D. Weinberger. NP completeness of Kauffman's NK model, a tuneable rugged fitness landscape (1996). Santa Fe Institute Working Paper 96-02-003, `http://www.santafe.edu/media/workingpapers/96-02-003.pdf`.

[10] S. Kirkpatrick and G. Toulouse. Configuration space analysis of travelling salesman problems. *Journal de Physique*, vol. 46 pp. 1277–1292 (1985).

[11] P. F. Stadler and W. Schnabl. The landscape of the traveling salesman problem. *Physics Letters A*, vol. 161 pp. 337–344 (1992).

[12] R. Karp. Reducibility among Combinatorial Problems. In R. Miller, J. Thatcher and J. Bohlinger (Editors), *Complexity of Computer Computations*, The IBM Research Symposia Series, pp. 85–103 (Springer US, 1972).

[13] R. A. Fisher. *The Genetical Theory of Natural Selection* (Clarendon Press, 1930).

[14] S. Wright. Evolution in Mendelian populations. *Genetics*, vol. 16 p. 97 (1931).

[15] M. Mitchell. *An introduction to genetic algorithms* (MIT Press, 1998).

[16] E. N. Gilbert. Gray Codes and Paths on the n-Cube. *Bell System Technical Journal*, vol. 37 pp. 815–826 (1958).

[17] F. Gray. Pulse code communication (1953). US Patent 2,632,058.

[18] P. Hegarty and A. Martinsson. On the existence of accessible paths in various models of fitness landscapes. *The Annals of Applied Probability*, vol. 24 pp. 1375–1395 (2014).

[19] L. Comtet. *Advanced combinatorics: The art of finite and infinite expansions.* (D. Reidel Publishing Co., Dordrecht, 1974), enlarged edn.

[20] Number of Permutations of [1..n] with k Components. *The On-Line Encyclopedia of Integer Sequences*. `https://oeis.org/A059438` (retrieved July 21, 2015).

[21] Number of paths (without loops) in graph of n-dimensional hypercube starting at point (0,0,0,...,0) and ending at (1,1,1,...,1). *The On-Line Encyclopedia of Integer Sequences*. `http://oeis.org/A059783` (retrieved July 21, 2015).

[22] J. Berestycki, É. Brunet and Z. Shi. Accessibility percolation with backsteps (2014). `arXiv:1401.6894`.

[23] J. F. C. Kingman. A Simple Model for the Balance between Selection and Mutation. *Journal of Applied Probability*, vol. 15 pp. pp. 1–12 (1978).

[24] S. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, vol. 128 pp. 11–45 (1987).

[25] B. Derrida. Random-energy model: Limit of a family of disordered models. *Physical Review Letters*, vol. 45 p. 79 (1980).

[26] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, vol. 24 pp. 2613–2626 (1981).

[27] T. Aita, H. Uchiyama, T. Inaoka, M. Nakajima, T. Kokubo and Y. Husimi. Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: Application to prolyl endopeptidase and thermolysin. *Biopolymers*, vol. 54 pp. 64–79 (2000).

[28] J. Franke, A. Klözer, J. De Visser and J. Krug. Evolutionary accessibility of mutational pathways. *PLoS Computational Biology*, vol. 7 p. e1002134 (2011).

[29] S. A. Kauffman and E. D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, vol. 141 pp. 211–245 (1989).

[30] P. Stadler. Landscapes and their correlation functions. *Journal of Mathematical Chemistry*, vol. 20 pp. 1–45 (1996).

[31] P. F. Stadler and R. Happel. Random field models for fitness landscapes. *Journal of Mathematical Biology*, vol. 38 pp. 435–478 (1999).

[32] J. Neidhart, I. G. Szendro and J. Krug. Exact results for amplitude spectra of fitness landscapes. *Journal of Theoretical Biology*, vol. 332 pp. 218–227 (2013).

[33] R. Heckendorn and D. Whitley. A Walsh analysis of NK landscapes. In T. Bäck (Editor), *Proceedings of the Seventh International Conference on Genetic algorithms*, pp. 41–48 (Morgan Kaufmann, 1997).

[34] D. M. Weinreich, Y. Lan, C. S. Wylie and R. B. Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current Opinion in Genetics & Development*, vol. 23 pp. 700–707 (2013).

[35] D. J. Gross and M. Mézard. The simplest spin glass. *Nuclear Physics B*, vol. 240 pp. 431–452 (1984).

[36] J. Buzas and J. Dinitz. An Analysis of Landscapes: Interaction Structure, Statistical Properties, and Expected Number of Local Optima. *IEEE Transactions on Evolutionary Computation*, vol. 18 pp. 807–818 (2014).

[37] S. Nowak and J. Krug. Analysis of adaptive walks on NK fitness landscapes with different interaction schemes. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2015 p. P06014 (2015).

[38] J. Neidhart, I. G. Szendro and J. Krug. Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model. *Genetics*, vol. 198 pp. 699–721 (2014).

[39] A. S. Perelson and C. A. Macken. Protein evolution on partially correlated landscapes. *Proceedings of the National Academy of Sciences*, vol. 92 pp. 9657–9661 (1995).

[40] B. Schmiegelt and J. Krug. Evolutionary Accessibility of Modular Fitness Landscapes. *Journal of Statistical Physics*, vol. 154 pp. 334–355 (2014).

[41] R. Durrett and V. Limic. Rigorous results for the NK model. *Annals of Probability*, vol. 31 pp. 1713–1753 (2003).

[42] E. D. Weinberger. Local properties of Kauffman's N-k model: A tunably rugged energy landscape. *Physical Review A*, vol. 44 pp. 6399–6413 (1991).

[43] S. N. Evans and D. Steinsaltz. Estimating Some Features of NK Fitness Landscapes. *The Annals of Applied Probability*, vol. 12 pp. pp. 1299–1321 (2002).

[44] D. M. Weinreich, N. F. Delaney, M. A. DePristo and D. L. Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, vol. 312 pp. 111–114 (2006).

[45] A. Martinsson. Unoriented first-passage percolation on the n-cube (2014). `arXiv:1402.2928`.

[46] A. Martinsson. Accessibility percolation and first-passage site percolation on the unoriented binary hypercube (2015). `arXiv:1501.02206`.

[47] S. Nowak and J. Krug. Accessibility percolation on n-trees. *EPL (Europhysics Letters)*, vol. 101 p. 66004 (2013).

[48] M. I. Roberts and L. Z. Zhao. Increasing paths in regular trees. *Electronic Communications in Probability*, vol. 18 pp. 1–10 (2013).

[49] X. Chen. *Branching random walks with selection*. Ph.D. thesis, Université Pierre et Marie Curie - Paris VI (2013). `https://tel.archives-ouvertes.fr/tel-00920308` (retrieved August 10, 2015).

[50] X. Chen. Increasing paths on N-ary trees (2014). `arXiv:1403.0843`.

[51] C. F. Coletti, R. J. Gava and P. M. Rodriguez. On the existence of accessibility in a tree-indexed percolation model (2014). `arXiv:1410.3320`.

[52] J. Berestycki, É. Brunet and Z. Shi. The number of accessible paths in the hypercube (2013). `arXiv:1304.0246`.

[53] M. Carneiro and D. L. Hartl. Adaptive landscapes and protein evolution. *Proceedings of the National Academy of Sciences*, vol. 107 pp. 1747–1751 (2010).

[54] D. Angluin and L. G. Valiant. Fast probabilistic algorithms for Hamiltonian circuits and matchings. In *Proceedings of the ninth annual ACM symposium on Theory of computing*, pp. 30–41 (ACM, 1977).

[55] L. Li. Phase transition for accessibility percolation on hypercubes (2015). `arXiv:1502.07642`.

[56] R. Ballerini and S. Resnick. Records from improving populations. *Journal of Applied probability*, pp. 487–502 (1985).

[57] J. Franke, G. Wergen and J. Krug. Records and sequences of records from random variables with a linear trend. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010 p. P10013 (2010).

[58] J. Franke and J. Krug. Evolutionary Accessibility in Tunably Rugged Fitness Landscapes. *Journal of Statistical Physics*, vol. 148 pp. 706–723 (2012).

[59] B. Schmiegelt and J. Krug. Sign epistasis and evolutionary accessibility in generalized NK-type fitness landscape models (2015+). In preparation.

[60] J. H. Gillespie. A simple stochastic gene substitution model. *Theoretical Population Biology*, vol. 23 pp. 202–215 (1983).

[61] J. H. Gillespie. Some Properties of Finite Populations Experiencing Strong Selection and Weak Mutation. *The American Naturalist*, vol. 121 pp. pp. 691–708 (1983).

[62] H. A. Orr. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. *Evolution*, vol. 52 pp. pp. 935–949 (1998).

[63] H. A. Orr. The population genetics of adaptation: The adaptation of DNA sequences. *Evolution*, vol. 56 pp. 1317–1330 (2002).

[64] C. A. Macken and A. S. Perelson. Protein evolution on rugged landscapes. *Proceedings of the National Academy of Sciences*, vol. 86 pp. 6191–6195 (1989).

[65] C. A. Macken, P. S. Hagan and A. S. Perelson. Evolutionary Walks on Rugged Landscapes. *SIAM Journal on Applied Mathematics*, vol. 51 pp. 799–827 (1991).

[66] H. Flyvbjerg and B. Lautrup. Evolution in a rugged fitness landscape. *Phys. Rev. A*, vol. 46 pp. 6714–6723 (1992).

[67] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, vol. 47 p. 713 (1962).

[68] J. H. Gillespie. *The causes of molecular evolution* (Oxford University Press, 1991).

[69] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger and P. Schuster. RNA folding and combinatory landscapes. *Physical Review E*, vol. 47 p. 2083 (1993).

[70] H. A. Orr. A Minimum on the Mean Number of Steps Taken in Adaptive Walks. *Journal of Theoretical Biology*, vol. 220 pp. 241–247 (2003).

[71] L. Bussolari, P. Contucci, M. D. Esposti and C. Giardinà. Energy-decreasing dynamics in mean-field spin models. *Journal of Physics A: Mathematical and General*, vol. 36 p. 2413 (2003).

[72] P. Contucci, C. Giardinà, C. Giberti, F. Unguendoli and C. Vernia. Interpolating greedy and reluctant algorithms. *Optimization Methods and Software*, vol. 20 pp. 509–514 (2005).

[73] P. Contucci, C. Giardinà, C. Giberti and C. Vernia. Finding Minima in Complex Landscapes: Annealed, Greedy and Reluctant Algorithms. *Mathematical Models and Methods in Applied Sciences*, vol. 15 pp. 1349–1369 (2005).

[74] M. Valente. An NK-like model for complexity. *Journal of Evolutionary Economics*, vol. 24 pp. 107–134 (2014).

[75] L. De Haan and A. Ferreira. *Extreme Value Theory: An Introduction* (Springer Science & Business Media, 2007).

[76] J. Neidhart and J. Krug. Adaptive Walks and Extreme Value Theory. *Physical Review Letters*, vol. 107 p. 178102 (2011).

[77] K. Jain. Number of adaptive steps to a local fitness peak. *EPL (Europhysics Letters)*, vol. 96 p. 58006 (2011).

[78] K. Jain and S. Seetharaman. Multiple adaptive substitutions during evolution in novel environments. *Genetics*, vol. 189 pp. 1029–1043 (2011).

[79] J. Neidhart. *Fitness Landscapes, Adaptation and Sex on the Hypercube*. Ph.D. thesis, Universität zu Köln (2014). `http://kups.ub.uni-koeln.de/5878` (retrieved July 21, 2015).

[80] S.-C. Park, J. Neidhart and J. Krug. Greedy adaptive walks on a correlated fitness landscape (2015). `arXiv:1507.03511`.

[81] T. Ernst. A method for q-calculus. *Journal of Nonlinear Mathematical Physics*, vol. 10 pp. 487–525 (2003).

[82] S.-C. Park, I. G. Szendro, J. Neidhart and J. Krug. Phase transition in random adaptive walks on correlated fitness landscapes. *Physical Review E*, vol. 91 p. 042707 (2015).

[83] F. Zanini and R. A. Neher. FFPopSim: an efficient forward simulation package for the evolution of large populations. *Bioinformatics*, vol. 28 pp. 3332–3333 (2012).

[84] N. J. Butterfield. Bangiomorpha pubescens n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology*, vol. 26 pp. 386–404 (2000).

[85] M. W. Feldman, S. P. Otto and F. B. Christiansen. Population genetic perspectives on the evolution of recombination. *Annual Review of Genetics*, vol. 30 pp. 261–295 (1996).

[86] S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, vol. 3 pp. 252–261 (2002).

[87] J. A. G. de Visser and S. F. Elena. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nature Reviews Genetics*, vol. 8 pp. 139–149 (2007).

[88] S. P. Otto. The evolutionary enigma of sex. *The American Naturalist*, vol. 174 pp. S1–S14 (2009).

[89] J. Maynard Smith. *The evolution of sex* (Cambridge University Press, 1978).

[90] R. E. Michod and B. R. Levin. *The Evolution of Sex: An Examination of Current Ideas* (Sinauer Associates Inc, 1987).

[91] B. Charlesworth and D. Charlesworth. An experiment on recombination load in Drosophila melanogaster. *Genetical Research*, vol. 25 pp. 267–273 (1975).

[92] C. R. Haag and D. Roze. Genetic load in sexual and asexual diploids: segregation, dominance and genetic drift. *Genetics*, vol. 176 pp. 1663–1678 (2007).

[93] A. Weismann. *Essays on heredity and kindred biological subjects* (Oxford University Press, 1889).

[94] A. Burt. Perspective: Sex, recombination, and the efficacy of selection - Was Weismann right? *Evolution*, vol. 54 pp. 337–351 (2000).

[95] H. J. Muller. The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 1 pp. 2–9 (1964).

[96] J. Felsenstein. The evolutionary advantage of recombination. *Genetics*, vol. 78 pp. 737–756 (1974).

[97] A. S. Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature*, vol. 336 pp. 435–440 (1988).

[98] H. J. Muller. Some genetic aspects of sex. *American Naturalist*, pp. 118–138 (1932).

[99] W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genetical research*, vol. 8 pp. 269–294 (1966).

[100] A. Altland, A. Fischer, J. Krug and I. G. Szendro. Rare events in population genetics: stochastic tunneling in a two-locus model with recombination. *Physical Review Letters*, vol. 106 p. 088101 (2011).

[101] L. Van Valen. A new evolutionary law. *Evolutionary Theory*, vol. 1 pp. 1–30 (173).

[102] L. Carroll. *Through the Looking-Glass, and What Alice Found There* (Macmillan Publishers Ltd, London, 1871).

[103] S. Nowak, J. Neidhart, I. G. Szendro and J. Krug. Multidimensional Epistasis and the Transitory Advantage of Sex. *PLoS Computational Biology*, vol. 10 p. e1003836 (2014).

[104] C. Rueffler, T. J. Van Dooren, O. Leimar and P. A. Abrams. Disruptive selection and then what? *Trends in Ecology & Evolution*, vol. 21 pp. 238–245 (2006).

[105] R. A. Watson, D. M. Weinreich and J. Wakeley. Genome structure and the benefit of sex. *Evolution*, vol. 65 pp. 523–536 (2011).

[106] P. Baldi and Y. Rinott. Asymptotic normality of some graph-related statistics. *Journal of Applied Probability*, pp. 171–175 (1989).

[107] L. Deecke. *Fitness landscapes and evolutionary accessibility: The effect of downhill steps*. Bachelor Thesis, Universität zu Köln (2015).

[108] É. Brunet. Private Communication (2015).

[109] R. Bürger. Linkage and the maintenance of heritable variation by mutation-selection balance. *Genetics*, vol. 121 pp. 175–184 (1989).

[110] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics* (Clarendon Press, Oxford, 1999).

## Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Joachim Krug betreut worden.

(Stefan Nowak)

## Teilpublikationen

- S. Nowak and J. Krug. Accessibility percolation on n-trees. *EPL (Europhysics Letters)*, vol. 101 p. 66004 (2013).
- S. Nowak, J. Neidhart, I. G. Szendro and J. Krug. Multidimensional Epistasis and the Transitory Advantage of Sex. *PLoS Computational Biology*, vol. 10 p. e1003836 (2014).
- S. Nowak and J. Krug. Analysis of adaptive walks on NK fitness landscapes with different interaction schemes. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2015 p. P06014 (2015).