Models for
Mutation, Selection, and Recombination
in Infinite Populations

Inauguraldissertation
zur
Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

an der Mathematisch-Naturwissenschaftlichen Fakultät
der
Ernst-Moritz-Arndt-Universität Greifswald

vorgelegt von
Oliver Redner,
geboren am 19.11.1972
in Hannover

Greifswald, den 26. März 2003

Dekan:	Prof. Dr. Jan-Peter Hildebrandt
1. Gutachter:	Prof. Dr. Michael Baake
2. Gutachter:	Prof. Dr. Reinhard Bürger
Tag der Promotion:	12. Mai 2003

Contents

Introduction

Biological evolution proceeds by the joint action of several elementary processes, the most important ones being mutation, selection, recombination, migration, and genetic drift. Their interaction is extremely complex and, as such, inaccessible to a mathematical treatment. For the latter, one therefore has to narrow the focus to isolated combinations of evolutionary objects and processes.

This thesis is concerned with population genetics, i.e., the study of the genetic structure of populations. We will consider two classes of models, which, together, comprise three of the most important evolutionary factors. The first one focuses on mutation and selection acting on a population of haploid, asexually reproducing individuals. The second class takes a somewhat complementary point of view by modeling an aspect of recombination known as unequal crossover. This is one possible cause for gene duplication, e.g., in rDNA sequences, and thus for the generation of redundancy on which mutation can act to produce evolutionary innovation. In both cases, environmental and developmental influences are neglected, and the individuals are taken to be fully described by their genotypes, possibly only with respect to a single character or trait.

From a mathematical perspective, the aim of this thesis is not primarily the advancement of a mathematical field but rather the fruitful application of those well-developed theories that are needed to tackle the biological problems in question. Among these are real, complex, and functional analysis, probability theory, as well as the theory of differential equations. The latter comes into play through the general assumption of an effectively infinite population size, that is, the exclusion of random genetic drift as an additional evolutionary factor. This allows for a deterministic formulation of the dynamics in terms of differential equations rather than by (stochastic) branching processes—which are a natural description if one deals with finite populations, and are only considered for conceptual purposes here. Furthermore, this thesis exclusively deals with the equilibrium behavior of the above models, which is described by eigenvalue equations of linear, respectively quadratic operators.

The outline is as follows. Chapters I and II are concerned with mutation–selection models. In this framework, selection is understood as the enhanced reproduction of fitter individuals at the cost of the less fit, where fitness is solely determined by the genotypes. Mutation is a random change of type, which may be modeled either as taking place during reproduction or as an independent process, going on in parallel. For a review and a guide to the vast body of literature on the subject, see [Bür00, Baa00].

Certain models for coupled mutation and selection, in which genotypes are taken to be sequences of fixed length and time proceeds in discrete steps, are formally equivalent to a model of statistical physics, the two-dimensional Ising model [Leu86, Leu87]. However, due to the complexity of the formalism employed to solve these, this relationship has led to few new results, e.g., [Tar92, Fra97]. Quite recently, the corresponding models with parallel mutation and selection in continuous time were observed to be analogous to the one-dimensional quantum version of the Ising model [Baa97, Baa98, Her01]. For the latter, rigorous solutions exist, which were applied to obtain expressions for the equilibrium values of the main quantities of biological importance, namely the population mean and variance of fitness and of the number of mutations in a sequence. The basis is a simple maximum principle for the mean fitness, which corresponds to the minimum principle of the free energy in statistical physics.

We have been able to generalize these results to models in which the genotypes are taken from any large but finite set, and to more general mutation and fitness schemes. The latter include quite diverse examples, ranging from a simple linear or quadratic dependence of fitness on the genotype over smoothly varying genotype–fitness mappings, such as those studied in [Cha90], to ones with sharp jumps, as in [Kon88, Eig89]. Furthermore, all derivations remain within classical probability theory and analysis, without reference to physics.

As an application, criteria could be given that determine whether a model exhibits discontinuous changes in its equilibrium behavior when the mutation rate is varied. Such phenomena have become known as error thresholds [Eig71]. The characterization of models exhibiting such thresholds has been a long-standing problem, see, e.g., [Swe82, Wie97]. For a more biologically interested audience, these results have been published in [Her02], including an appendix describing the connection to physics. In Chapter I, they are put together in a rigorous and condensed form for a mathematical readership.

Afterwards, Chapter II turns to another important class of mutation–selection models, the so-called continuum-of-alleles (COA) models, in which genotypes are taken from a continuous set. These pay respect to the assumption that at a gene locus effectively infinitely many alleles can be generated and every mutation results in a new allele, cf. [Kim65]. The first part of the chapter relates the COA model to models with discrete genotypes in describing an approximation procedure. Mathematically, this is a generalization of standard methods of approximation theory to special cases of non-compact operators. This treatment is necessary for the justification of numerical analysis, which is inevitably discrete in nature, and it allows the transfer of results. In a second part, first steps are taken towards a simple maximum principle for the mean fitness by generalizing our findings for the models with discrete genotypes.

Finally, Chapter III is devoted to unequal crossover models recently introduced in [Shp02] as modifications of previous models [Oht83, Wal87]. One considers the size evolution of sequences that contain repeated units when the alignment of two recombining sequences is possibly imperfect. This leads to a redistribution of the building blocks among the participating sequences. For some of the conjectures in [Shp02], we are able to give rigorous proofs, mainly concerning the convergence of the distribution of an infinite population towards the known equilibria.

Throughout this thesis, the following notation will be used. Vectors and matrices are denoted by bold symbols, e.g., $\boldsymbol{p}$ and $\boldsymbol{M}$. Their components are referred to as, e.g., p_i and M_{ij}, respectively. At some points, references to sections, equations, propositions, etc. are necessary across chapter boundaries. In these cases, the roman chapter number is prepended, e.g., Section I.3.4 and (I.30).

I

Mutation–selection models

This chapter and the following are concerned with models for mutation and selection, in which effects of other evolutionary forces are neglected. These models are introduced in Section 1 on a general basis. Afterwards, for the rest of this chapter, we restrict ourselves to the case that individuals are characterized by genotypes from a large but finite set. Section 2 introduces the specific model under consideration and connects it to a multitype branching process, both forward and backward in time. The latter direction gives rise to the definition of the distribution of ancestors, which will play an important role in the sequel. In Section 3, the main results, which allow for a simple characterization of the equilibrium population, are formulated, proved, and discussed. As an application, Section 4 treats threshold phenomena that may occur when the mutation rates are varied, such as the well-known error threshold. Chapter II is then devoted to so-called continuum-of-alleles (COA) models, in which the genotypes are characterized by the elements of a continuous set, such as $\mathbb{R}$ or the interval $[0,1]$. As a general reference for mutation–selection models, Bürger's book [Bür00] is recommended.

1 The general model

We consider the evolution of an effectively infinite population of haploid[1] individuals subject to mutation and selection. Disregarding environmental effects, we take individuals to be fully described by their genotypes, which are labeled by the elements of some set Γ endowed with a positive σ-finite measure ν (usually the Haar measure). This set may either be finite, $\Gamma = \{1, \dots, M\}$, with ν being the counting measure, or an interval $\Gamma \subset \mathbb{R}$ equipped with the Lebesgue measure. The set Γ may be taken either as the whole genome, or as the genomic basis of a specific trait or function (i.e., an observable phenotypical property). We will describe the population at time t by a probability density on Γ, i.e., an integrable function $p(t) \in \mathrm{L}^1(\Gamma, \nu)$ with $p(t) \geq 0$ and $\int_\Gamma p(x,t)\,\mathrm{d}\nu(x) = 1$.[2]

Throughout this and the following chapter, we will use the formalism for overlapping generations, which works in continuous time, and only comment on extensions to the analogous model for subsequent generations in discrete time. The standard equation that describes the evolution of the density $p(t)$ is, cf. [Kim65] and [Bür00, (IV.1.3)],

$$\dot{p}(x,t) = \big(r(x) - \bar{r}(t)\big)p(x,t) + \int_\Gamma \big(u(x,y)\,p(y,t) - u(y,x)\,p(x,t)\big)\,\mathrm{d}\nu(y)\,. \tag{1}$$

[1] Diploid individuals without dominance may be described by the same formalism since the reproduction rate of an allele combination is additive with respect to both alleles, compare [Bür00, Sec. III.2.1].

[2] For a treatment of the general case of arbitrary locally compact spaces Γ and the description in terms of probability measures, see [Bür00, Sec. IV].

Here, $r(x)$ is the Malthusian fitness, or (effective) reproduction rate, of type $x \in \Gamma$, which is connected to the respective birth and death rates as $r(x) = b(x) - d(x)$, and $\bar{r}(t) = \int_\Gamma r(x)\, p(x,t)\, \mathrm{d}\nu(x)$ designates the mean fitness. The mutation rates and the distributions of mutant types are given by u, where $u(x,y)$ corresponds to a mutation from type y to x, and the dot denotes the time derivative $\partial/\partial t$.

In this model, mutation and selection are assumed to be independent processes, going on in parallel. However, mutation may also be viewed as occuring during reproduction. In this case, we have $u(x,y) = v(x,y)\, b(y)$, where $v(x,y)$ gives the respective mutation *probability* during a reproduction event and the distribution of mutants. Since, formally, this leads to the same type of model, it will not be discussed separately.

2 The model with discrete genotypes

2.1 Deterministic description

Let us, for the rest of this chapter, turn exclusively to the model with discrete genotypes, for which $\Gamma = \{1, \ldots, M\}$. Here, we identify the population density $p(t)$ with a vector $\boldsymbol{p}(t) = (p_i(t))_{1 \le i \le M}$, which reflects the canonical coordinatization. The evolution equation (1) becomes the following system of ordinary differential equations, cf. [Cro70, Hof85],

$$\dot{p}_i(t) = \big(R_i - \bar{R}(t)\big) p_i(t) + \sum_j \big(m_{ij}\, p_j(t) - m_{ji}\, p_i(t)\big)\,. \qquad (2)$$

For reasons that will become clear in Section 2.6, we use capital letters for the reproduction rates R_i here. Further, we write m_{ij} for the mutation rate from type j to i.[3]

For some of the main results of this chapter, further assumptions on the mutation scheme are required. To this end, we collect genotypes into classes $\mathcal{X}_k$ of equal fitness, $0 \le k \le N$, and assume mutations only to occur between neighboring classes. Let R_k denote the fitness of class k and $U_k^{\pm}$ the mutation rate from class $\mathcal{X}_k$ to $\mathcal{X}_{k\pm1}$ (i.e., the total rate for each genotype in $\mathcal{X}_k$ to mutate to some genotype in $\mathcal{X}_{k\pm1}$), with the convention $U_0^- = U_N^+ = 0$. Thus, we obtain a variant of the so-called single-step mutation model,

$$\dot{p}_k(t) = \big(R_k - \bar{R}(t) - U_k^+ - U_k^-\big) p_k(t) + U_{k-1}^+\, p_{k-1}(t) + U_{k+1}^-\, p_{k+1}(t)\,. \qquad (3)$$

(Here, the convention $p_{-1}(t) = p_{N+1}(t) = 0$ is used.) We can, for example, think of $\mathcal{X}_0$ as the (mutation-free) wildtype class with maximum fitness and fitness only depending on the number of mutations carried by an individual. If, further, mutation is modeled as a continuous process (or if multiple mutations during reproduction can be ignored), Equation (2) reduces to (3), with an appropriate choice of mutation classes. Depending on the realization one has in mind, the U_k then describe the total mutation rate affecting the whole genome or just the trait or function under consideration.

In most of our examples, we will use the Hamming graph as our genotype space. Here, genotypes are represented as binary sequences $\boldsymbol{s} = s_1 s_2 \ldots s_N \in \{+,-\}^N$, hence $M = 2^N$.

[3] Generally throughout this thesis, I use this index convention, which is the transposed of what is common for Markov processes but more natural in the context of differential equations.

$$+ \;\xrightleftharpoons[(1-\kappa)\mu]{(1+\kappa)\mu}\; -$$

Figure 1: Rates for mutations and back mutations at each site or locus of a biallelic sequence.

The two possible values at each site, $+$ and $-$, may be understood either in a molecular context as nucleotides (purines and pyrimidines) or, on a coarser level, as wildtype and mutant alleles of a biallelic multilocus model. We will assume equal mutation rates at all sites, but allow for different rates, $(1+\kappa)\mu$ and $(1-\kappa)\mu$, for mutations from $+$ to $-$ and for back mutations, respectively, according to the scheme depicted in Figure 1. Here, $\mu \geq 0$ describes the overall mutation rate and $\kappa \in [-1, 1]$ is an asymmetry parameter.

Clearly, the biallelic model reduces to a single-step mutation model (with the same N) if the fitness landscape[4] is invariant under permutation of sites. To this end, we distinguish a reference genotype $\boldsymbol{s}_+ = ++\ldots+$, in most cases the wildtype, and assume that the fitness $R_{\boldsymbol{s}}$ of sequence $\boldsymbol{s}$ depends only on the Hamming distance $k = d_{\mathrm{H}}(\boldsymbol{s}, \boldsymbol{s}_+)$ to $\boldsymbol{s}_+$ (i.e., the number of mutations, or '$-$' signs in the sequence). The resulting total mutation rates between the Hamming classes $\mathcal{X}_k$ and $\mathcal{X}_{k\pm1}$ read

$$U_k^+ = (1+\kappa)\,\mu\,(N-k) \quad \text{and} \quad U_k^- = (1-\kappa)\,\mu\,k \tag{4}$$

if mutation is assumed to be an independent process at all sites. We usually have the situation in mind in which fitness decreases with k and will therefore speak of U_k^+ and U_k^- as the *deleterious* and *advantageous* mutation rates. However, monotonic fitness is never *assumed*, unless this is stated explicitly.

In much of the following, we will treat the general model (2), which builds on single genotypes, and the single-step mutation model (3), in which the units are genotype classes, with the help of a common formalism. To this end, note that both models can be recast into the following general form using matrices of dimension M, respectively $N+1$:

$$\dot{\boldsymbol{p}}(t) = \big(\boldsymbol{H} - \bar{R}(t)\mathbb{1}\big)\boldsymbol{p}(t)\,. \tag{5}$$

Here, $\mathbb{1}$ is the identity. The matrix $\boldsymbol{H} = \boldsymbol{R} + \boldsymbol{M}$ is composed of a diagonal matrix $\boldsymbol{R}$ that holds the Malthusian fitness values, and the mutation matrix $\boldsymbol{M} = (M_{ij})$ with either off-diagonal entries reading m_{ij}, or with $U_k^\pm$ on the secondary diagonals. The diagonal elements in each case are $M_{ii} = -\sum_{j\neq i} M_{ji}$, hence the column sums vanish, i.e., $\boldsymbol{M}$ is a Markov generator. Where the more restrictive form of the single-step model is needed, this will be stated explicitly. Unless we talk about unidirectional mutation ($U_k^- \equiv 0$ for the single-step mutation model), we will always assume that $\boldsymbol{M}$ is irreducible (i.e., each entry is non-zero for a suitable power of $\boldsymbol{M}$).

Let now $\boldsymbol{T}(t) := \exp(t\boldsymbol{H})$, with matrix elements $T_{ij}(t)$. Then, the solution of (5) is given by (see, e.g., [Bür00, Sec. III.1])

$$\boldsymbol{p}(t) = \frac{\boldsymbol{T}(t)\boldsymbol{p}(0)}{\sum_{i,j} T_{ij}(t) p_j(0)}\,, \tag{6}$$

[4] We use the notion of a *fitness landscape* [Kau87] as synonymous with *fitness function* for the mapping from genotypes to individual fitness values.

as can easily be established by using $\sum_{i,j} H_{ij} p_j(t) = \sum_i R_i p_i(t) = \bar{R}(t)$ and differentiating.[5] Due to the irreducibility, the population vector converges to a unique, globally stable equilibrium distribution $\boldsymbol{p} := \lim_{t\to\infty} \boldsymbol{p}(t)$ with $p_i > 0$ for all i, which describes mutation–selection balance (cf., e.g., [Bür00, Sec. IV.2]). By the Perron–Frobenius theorem (compare [Sch74, Sec. I.6] or [Gan86, Sec. 13.2]), $\boldsymbol{p}$ is the (right) eigenvector corresponding to the largest eigenvalue, $\lambda_{\max}$, of $\boldsymbol{H}$. (Strictly speaking, if there are negative fitness values, we have to add a suitable constant C to all fitness values to make $\boldsymbol{H}$ positive, i.e., all its entries non-negative. Then, $\lambda_{\max}$ is given by the spectral radius of $\boldsymbol{H}+C$, which is its largest eigenvalue, minus C.) For unidirectional mutation, the equilibrium distribution $\boldsymbol{p}$ is in general not unique, see the discussion in Section 3.2.

2.2 The corresponding branching process

Our approach will heavily rely on genealogical relationships, which contain more detailed information than the time evolution of the relative frequencies (6) alone. On the next pages, we therefore reconsider the mutation–selection model as a branching process.[6]

We consider the process of mutation, reproduction and death as a (continuous-time) multitype branching process, as described previously for the discrete-time variant of the so-called quasispecies model, i.e., the analogous model with mutation coupled to reproduction [Dem85, Hof88, Ch. 11.5]. Let us start with a *finite* population of individuals, each described by an element i of a finite set $\mathit{\Gamma}$, that reproduce (at rates B_i), die (at rates D_i), or change type (at rates M_{ij}) independently of each other, without any restriction on population size. Let $Y_i(t)$ be the random variable denoting the number of individuals of type i at time t, and $n_i(t)$ the corresponding realization; collect the components into vectors $\boldsymbol{Y}$, $\boldsymbol{n} \in \mathbb{N}_0{}^{\mathit{\Gamma}}$, and let $\boldsymbol{e}_i$ be the i-th unit vector. The transition probabilities for the joint distribution, $\Pr\big(\boldsymbol{Y}(t) = \boldsymbol{n}(t) \mid \boldsymbol{Y}(0) = \boldsymbol{n}(0)\big)$, which we will abbreviate as $\Pr\big(\boldsymbol{n}(t) \mid \boldsymbol{n}(0)\big)$ by abuse of notation, are governed by the differential equation[7]

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \Pr\big(\boldsymbol{n}(t) \mid \boldsymbol{n}(0)\big) = -\,\Big(\sum_i (B_i + D_i + \sum_{j\neq i} M_{ji})\, n_i(t)\Big) \Pr\big(\boldsymbol{n}(t) \mid \boldsymbol{n}(0)\big) \\
+ \sum_i B_i \big(n_i(t) - 1\big) \Pr\big(\boldsymbol{n}(t) - \boldsymbol{e}_i \mid \boldsymbol{n}(0)\big) \\
+ \sum_i D_i \big(n_i(t) + 1\big) \Pr\big(\boldsymbol{n}(t) + \boldsymbol{e}_i \mid \boldsymbol{n}(0)\big) \\
+ \sum_{\substack{i,j \\ i\neq j}} M_{ij} \big(n_j(t) + 1\big) \Pr\big(\boldsymbol{n}(t) - \boldsymbol{e}_i + \boldsymbol{e}_j \mid \boldsymbol{n}(0)\big)\,.
\end{aligned}
\tag{7}
$$

[5]Alternatively, one may apply Thompson's trick [Tho74] and consider $\boldsymbol{q}(t) = \exp(\int_0^t \bar{R}(\tau)\,\mathrm{d}\tau)\boldsymbol{p}(t)$, for which the linear equation $\dot{\boldsymbol{q}}(t) = \boldsymbol{H}\boldsymbol{q}(t)$ and thus $\boldsymbol{q}(t) = \boldsymbol{T}(t)\boldsymbol{p}(0)$ holds.

[6]Whereas these considerations are crucial for a deeper understanding of the results of this chapter and make clear the terminology used, they are a detour from a technical point of view. The impatient reader may therefore directly proceed to Section 2.3, which contains a summary of the main quantities and notation introduced.

[7]Note that differentiability of the transition *probabilities* is guaranteed in a finite-state, continuous-time Markov chain, provided the transition *rates* are finite, cf. [Kar75, Ch. 4] and [Kar81, Ch. 14].

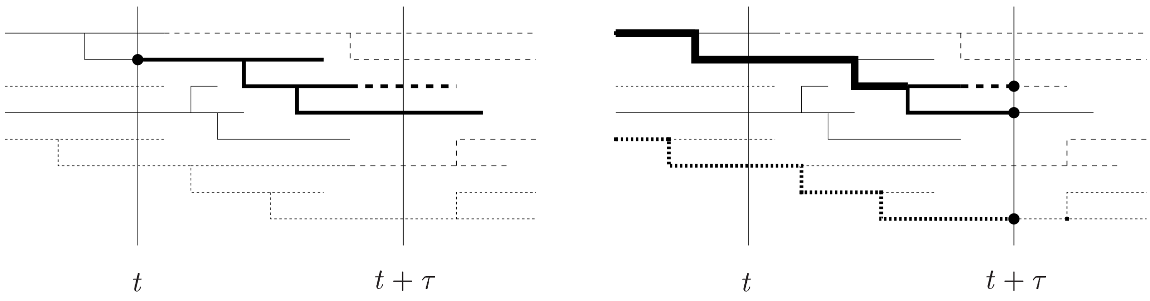

Figure 2: The multitype branching process. Individuals reproduce (branching lines), die (ending lines), or mutate (lines changing type) independently of each other; the various types are indicated by different line styles. Left: The fat lines mark the clone founded by a single individual (bullet) at time t. Right: The fat lines mark the lines of descent defined by three individuals (bullets) at time $t + \tau$. After coalescence of two lines, their ancestor receives twice the 'weight', as indicated by extra fat lines.

The connection of this stochastic process with the deterministic model described in Section 2.1 is twofold. Firstly, in the limit of an infinite number of individuals (i.e., $n := \sum_i n_i(0) \to \infty$), the sequence of random variables $\boldsymbol{Y}^{(n)}(t)/n$ converges almost surely to the solution $\boldsymbol{y}(t)$ of $\dot{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ with initial condition $\boldsymbol{y}(0) = \boldsymbol{n}(0)/n$ [Eth86, Thm. 11.2.1], $\Pr\big(\lim_{n\to\infty} \boldsymbol{Y}^{(n)}(t)/n = \boldsymbol{y}(t)\big) = 1$. (The superscript (n) denotes the dependence on the number of individuals.) The connection is now clear since $\boldsymbol{p}(t) := \boldsymbol{y}(t)/\sum_i y_i(t)$ solves the mutation–selection equation (2).

Secondly, taking expectations of Y_i and marginalizing over all other variables, one obtains the differential equation for the conditional expectations

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} E\big(Y_i(t) \mid \boldsymbol{n}(0)\big) &= (B_i - D_i) E\big(Y_i(t) \mid \boldsymbol{n}(0)\big) \\ &\quad + \sum_j \big[M_{ij} E\big(Y_j(t) \mid \boldsymbol{n}(0)\big) - M_{ji} E\big(Y_i(t) \mid \boldsymbol{n}(0)\big)\big]\,.\end{aligned} \tag{8}$$

Clearly, the matrix $\boldsymbol{H}$ appears as the (infinitesimal) generator here, and the solution is given by $\boldsymbol{T}(t)\,\boldsymbol{n}(0)$, where $\boldsymbol{T}(t) := \exp(t\boldsymbol{H})$ is the corresponding positive semigroup (see also [Hof88, Ch. 11.5]). In particular, we have $E\big(Y_i(t) \mid \boldsymbol{e}_j\big) = T_{ij}(t)$ for the expected number of i-individuals at time t, in a population started by a single j-individual at time 0 (a 'j-clone'). In the same way, $T_{ij}(\tau)$ is the expected number of descendants of type i at time $t + \tau$ in a j-clone started at an arbitrary time t, cf. the left panel of Figure 2. (Note that, due to the independence of individuals and the Markov property and homogeneity of the process on the 'large' state space ${\mathbb{N}_0}^{\Gamma}$, the progeny distribution depends only on the age of the clone, and on the founder type.) Further, the expected total size of a j-clone of age τ, irrespective of the descendants' types, is $\sum_i T_{ij}(\tau)$.

Initial conditions come into play if we consider the reproductive success of a clone relative to the whole population. A population of independent individuals, with initial composition $\boldsymbol{p}(t)$, has expected mean clone size $\sum_{i,j} T_{ij}(\tau) p_j(t)$ at time $t + \tau$ (here, t always means 'absolute' time, whereas τ denotes a time increment). The expected size of a single j-clone at time $t + \tau$, relative to the expected mean clone size of the whole population, then is

$$z_j(\tau,t) := \sum_i T_{ij}(\tau) / \sum_{k,\ell} T_{k\ell}(\tau)\, p_\ell(t)\,. \tag{9}$$

The z_j express the expected relative success of a type after evolution for a time interval τ in the sense that, if $z_j(\tau,t) > 1$ (< 1), we can expect the clone to flourish more (less) than average (this does in general not mean that type j is expected to increase (decrease) in abundance relative to the initial population). Clearly, the values of the z_j depend on the fitness of type j, but also on its mutation rate and the fitness of its (mutated) offspring. (If there is only mutation, but no reproduction or death, one has a Markov chain even on the 'small' state space Γ and $z_j(\tau,t) \equiv 1$.)

We now consider lines of descent, as in the right panel of Figure 2. To this end, we randomly pick an individual alive at time $t+\tau$, and trace its ancestry back in time; this results in an unbranched line (in contrast to the lineage forward in time). Let $Z_{t+\tau}(t)$ denote the type found at time $t \leq t+\tau$, where we will drop the index for easier readability. We seek its probability distribution $\Pr\big(Z(t)=j\big)$. Since the (relative) clone size $z_j(\tau,t)$ also determines the expected (relative) frequency of lines present at time $t+\tau$ that contain a j-type ancestor at time t, we have

$$\Pr\big(Z(t)=j\big) = z_j(\tau,t)\, p_j(t) =: a_j(\tau,t)\,. \tag{10}$$

The $a_j(\tau,t)$ define a probability distribution ($\sum_j a_j(\tau,t) \equiv 1$), which will be of major importance, and may be interpreted in two ways. Forward in time, $a_j(\tau,t)$ is the frequency of j-individuals at time t, weighted by their relative number of descendants after evolution for some time τ. Looking backward in time, $a_j(\tau,t)$ is the fraction of the ($\boldsymbol{p}$-distributed) population at time $t+\tau$ whose ancestor at time t is of type j. We shall therefore refer to $\boldsymbol{a}(\tau,t)$ as the *ancestral distribution* at the earlier time, t.

Let us, at this point, expand a little further on this backward picture by explicitly constructing the time-reversed process. This is done in the usual way, by writing the joint distribution of parent–offspring pairs (i.e., pairs $Z(t)$ and $Z(t+\tau)$) in terms of forward and backward transition probabilities. On the one hand,

$$\begin{aligned}\Pr\big(Z(t+\tau)=i, Z(t)=j\big) &= \Pr\big(Z(t+\tau)=i \mid Z(t)=j\big)\Pr\big(Z(t)=j\big)\\ &= P_{ij}(\tau)\, a_j(\tau,t)\,.\end{aligned} \tag{11}$$

Here, the $P_{ij}(\tau) := \Pr\big(Z(t+\tau)=i \mid Z(t)=j\big)$ may be obtained by rewriting the (conditional) expectations defining the (forward) branching process as $T_{ij}(\tau) = P_{ij}(\tau)\sum_k T_{kj}(\tau)$, which gives

$$P_{ij}(\tau) = T_{ij}(\tau) / \sum_k T_{kj}(\tau). \tag{12}$$

On the other hand,

$$\begin{aligned}\Pr\big(Z(t+\tau)=i, Z(t)=j\big) &= \Pr\big(Z(t)=j \mid Z(t+\tau)=i\big)\Pr\big(Z(t+\tau)=i\big)\\ &= \tilde{P}_{ji}(\tau,t)\, p_i(t+\tau)\,,\end{aligned} \tag{13}$$

where $\tilde{P}_{ji}(\tau,t) := \Pr\big(Z(t)=j \mid Z(t+\tau)=i\big)$ is the transition probability of the time-reversed process and is obtained as $\tilde{P}_{ji}(\tau,t) = a_j(\tau,t)P_{ij}(\tau)\big(p_i(t+\tau)\big)^{-1}$ from (11) and

(13). With Equations (9), (10), and (12), one therefore obtains the elements of the backward transition matrix $\tilde{\boldsymbol{P}}$ as

$$\tilde{P}_{ji}(\tau,t) = p_j(t)\frac{T_{ij}(\tau)}{\sum_{k,\ell} T_{k\ell}(\tau)\, p_\ell(t)}\big(p_i(t+\tau)\big)^{-1}. \tag{14}$$

By differentiating $\tilde{\boldsymbol{P}}(\tau,t)$ with respect to τ and evaluating it at $\tau = 0$, one obtains the matrix $\boldsymbol{Q}(t)$ governing the corresponding backward process in continuous time. Its elements read $Q_{ji}(t) = \frac{d}{d\tau}\tilde{P}_{ji}(\tau,t)\big|_{\tau=0} = p_j(t)\big(H_{ij} - \delta_{ij}\bar{R}(t)\big)\big(p_i(t)\big)^{-1} - \delta_{ij}\dot{p}_i(t)/p_i(t)$. Using (5) this simplifies to

$$Q_{ji}(t) = \begin{cases} p_j(t)H_{ij}\big(p_i(t)\big)^{-1} & \text{for } i \neq j, \\ -\sum_{k\neq i} p_k(t)H_{ik}\big(p_i(t)\big)^{-1} & \text{for } i = j. \end{cases} \tag{15}$$

Note that the backward process is, in general, state-dependent (it does not represent a Markov chain). Note also that time reversal works in the same way if sets of types $\mathcal{X}_k$ instead of single types are considered, as long as mutation and reproduction rates are the same within classes. Furthermore, an analogous treatment is possible both for mutation coupled to reproduction, as well as for subsequent generations.

As to the asymptotic behavior of our branching process, it is well-known that, for irreducible $\boldsymbol{H}$ and $t \to \infty$, the time evolution matrix $\exp\big(t(\boldsymbol{H} - \lambda_{\max}\mathbb{1})\big)$ becomes a projector onto the equilibrium distribution $\boldsymbol{p}$, with matrix elements $p_i z_j$ (e.g., [Kar81, App.]). Here, $\boldsymbol{z}$ is the Perron–Frobenius (PF) left eigenvector of $\boldsymbol{H}$, normalized such that $\sum_i z_i p_i = 1$. As suggested by our notation, one also has

$$\lim_{t,\tau\to\infty} \boldsymbol{z}(\tau,t) = \boldsymbol{z}\,, \tag{16}$$

which follows from (9).[8] We therefore term z_i the *relative reproductive success* of type i.

The *stationary* backward process is governed by the matrix $Q_{ji} = p_j\big(H_{ij} - \delta_{ij}\lambda_{\max}\big)p_i^{-1}$, which can now be interpreted as a Markov generator. Further, the (asymptotic) ancestral distribution, given by $a_i = z_i p_i$, turns out to be the equilibrium distribution of the backward process, since $\sum_i Q_{ji}a_i = \sum_i p_j(H_{ij} - \delta_{ij}\lambda_{\max})p_i^{-1}z_i p_i = \sum_i p_j z_i(H_{ij} - \delta_{ij}\lambda_{\max}) = 0$. Due to ergodicity of the latter ($\boldsymbol{Q}$ is irreducible if $\boldsymbol{H}$ is), $\boldsymbol{a}$ is, at the same time, the distribution of types along each line of descent (with probability 1).

2.3 The equilibrium ancestral distribution

As we saw in Section 2.2, there is a simple link between the algebraic properties of $\boldsymbol{H}$ and the probabilistic structure of the mutation–selection process at equilibrium, which

[8] Both $\boldsymbol{z}$ and $\boldsymbol{p}$ also admit a more stochastic interpretation. If the population does not go to extinction, one has $\lim_{t\to\infty} Y_i(t)/\sum_j Y_j(t) = p_i$ almost surely, see [Ath72, Thm. V.7.2], which is the continuous-time analog of the Kesten–Stigum theorem for discrete time [Kes66, Kur97]. Further, for the *critical process* generated by $\boldsymbol{H} - \lambda_{\max}\mathbb{1}$, one has $\lim_{t\to\infty} t\,\Pr\big(\boldsymbol{Y}(t) \neq \boldsymbol{0}\,|\,\boldsymbol{Y}(0) = \boldsymbol{e}_j\big) = z_j/C$, where C is a constant, and $\lim_{t\to\infty} \frac{1}{t}E\big(Y_i(t)\mid \boldsymbol{Y}(0) = \boldsymbol{e}_j, \boldsymbol{Y}(t) \neq \boldsymbol{0}\big) = Cp_i$; this is the continuous-time analog of a result by Jagers [Jag75, p. 94]. Note that, in the long run, the expected number of offspring depends on the founder type only through the probability of nonextinction of its progeny.

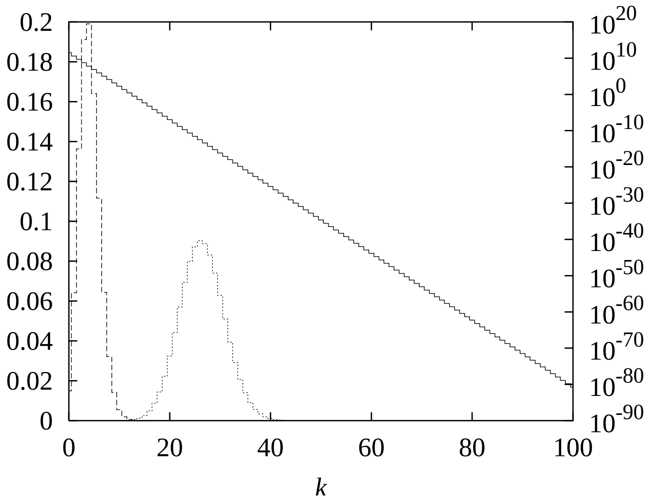

Figure 3: Equilibrium values of population frequencies p_k (dotted line), ancestral frequencies a_k (dashed line), and relative reproductive success z_k (solid line) for the biallelic model with additive fitness $R_k = \gamma(N-k)$ (where γ is the loss in reproduction rate due to a single mutation), point mutation rate $\mu = 0.2\gamma$, mutation asymmetry parameter $\kappa = \frac{1}{2}$, and sequence length $N = 100$. The logarithmic right axis refers to the z_k only.

may be summarized as follows. The PF right eigenvector $\boldsymbol{p}$ (with $\sum_i p_i = 1$) determines the composition of the population at mutation–selection balance; the corresponding left eigenvector $\boldsymbol{z}$ (normalized so that $\sum_i z_i p_i = 1$) contains the asymptotic offspring expectation (or relative reproductive success) of the various types; and the ancestral distribution, defined by $a_i = p_i z_i$, gives the asymptotic distribution of types that are met when lines of descent are followed backward in time (cf. Figure 2). Figure 3 shows $\boldsymbol{p}$, $\boldsymbol{a}$, and $\boldsymbol{z}$ for a single-step mutation model with linear fitness. One sees that z_k decreases exponentially.

For the single-step mutation model, we may directly transform the eigenvalue equation $\boldsymbol{Hp} = \lambda_{\max}\boldsymbol{p}$ into an equation for $\boldsymbol{a}$. To this end, we define a diagonal transformation matrix $\boldsymbol{S}$ with non-zero elements $S_{kk} = \prod_{\ell=1}^{k} \sqrt{U_\ell^- / U_{\ell-1}^+}$ and obtain a symmetric matrix by $\tilde{\boldsymbol{H}} := \boldsymbol{SHS}^{-1}$. The corresponding PF right and left eigenvectors are given by $\tilde{\boldsymbol{p}} = \boldsymbol{Sp}$ and $\tilde{\boldsymbol{z}} = \boldsymbol{S}^{-1}\boldsymbol{z}$. But now, as $\tilde{\boldsymbol{H}}$ is symmetric, we have $\tilde{\boldsymbol{z}} \sim \tilde{\boldsymbol{p}}$ (where $\sim$ means proportional to). Hence, due to $a_k = z_k p_k = \tilde{z}_k \tilde{p}_k \sim \tilde{p}_k^2$, one has $\tilde{p}_k \sim \sqrt{a_k}$. Thus, we obtain the following explicit form of the eigenvalue equation for $\tilde{\boldsymbol{H}}$:

$$\left(R_k - U_k^+ - U_k^-\right)\sqrt{a_k} + \sqrt{U_{k-1}^+ U_k^-}\sqrt{a_{k-1}} + \sqrt{U_k^+ U_{k+1}^-}\sqrt{a_{k+1}} = \lambda_{\max}\sqrt{a_k}\,. \tag{17}$$

Note that (17) relates the mean fitness of the *equilibrium* population ($\bar{R} = \lambda_{\max}$) to the *ancestral* frequencies a_k.

This property of the single-step mutation model, that there is a *diagonal* transformation matrix which symmetrizes the matrix $\boldsymbol{H}$ and makes it interpretable in terms of the ancestral distribution, will be crucial for the derivation of the main results of this chapter. In the next chapter, it will be generalized to continuous genotype spaces, for which it will find a similarly useful application.

2.4 Observables and averages

Now, we define the observables, i.e., quantities that can (in principle) be measured, which are used to describe the population. Besides the usual population mean, we shall also introduce the mean with respect to the ancestral distribution (see the Section 2.3).

We will consider means and variances of two observables. These are, for each type (or class) i, its *fitness value* R_i and its *mutational distance* X_i from the reference genotype (or the class $\mathcal{X}_0$). For the biallelic model in particular, mutational distance corresponds to the Hamming distance to $\boldsymbol{s}_+$. If, in addition, this is the fittest type, X_i just gives the number of deleterious mutations. But in general it can also be used to describe the value of any additive trait with equal contributions of sites or loci. Similarly, for single-step mutation, we define X_k to be the distance from the class $\mathcal{X}_0$, thus $X_k = k$ for class $\mathcal{X}_k$. Again, X_k may be viewed as (the genetic contribution to) any character with discrete values that depends linearly on the mutation classes.

Representing an arbitrary observable as (O_i), such as (R_i) or (X_i), we will denote its *population average* as

$$\bar{O}(t) := \sum_i O_i \, p_i(t) \,. \tag{18}$$

By omission of the time dependence we will indicate the equilibrium average (with respect to the unique distribution $\boldsymbol{p}$, cf. the end of Section 2.1).

As to mean fitness, $\bar{R}(t)$ determines the *mutation load*, $L(t) := R_{\max} - \bar{R}(t)$. Here, $R_{\max} = \max_i R_i$ is the fitness of the fittest genotype, in line with the usual convention (see, e.g., [Ewe79, Bür00]). It is well-known that the equilibrium value $\bar{R} := \lim_{t\to\infty} \bar{R}(t)$ is given by the largest eigenvalue, $\lambda_{\max}$, of $\boldsymbol{H}$.

For the variance of fitness, $V_R(t) = \sum_i (R_i - \bar{R}(t))^2 \, p_i(t)$, we differentiate $\bar{R}(t)$ according to (2), i.e., $\frac{\mathrm{d}}{\mathrm{d}t}\bar{R}(t) = \sum_i R_i \dot{p}_i(t) = V_R(t) + \sum_{i,j} R_i M_{ij} p_j(t)$, and hence

$$V_R(t) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{R}(t) - \sum_{i,j} R_i M_{ij} p_j(t) = \frac{\mathrm{d}}{\mathrm{d}t}\bar{R}(t) + \sum_j \Big(\sum_i (R_j - R_i) M_{ij} \Big) p_j(t) \,. \tag{19}$$

The interpretation of this completely general formula is as follows: In absence of mutation, (19) just reproduces Fisher's Fundamental Theorem, i.e., the variance of fitness equals the change in mean fitness, as long as there is no dominance (see, e.g., [Ewe79]). If mutation is present, however, a second component emerges, which is given by the population mean of the mutational effects on fitness (see below for a definition), weighted by the corresponding rates. It may be understood as the *rate of change in mean fitness due to mutation alone*. At mutation–selection balance, this second term is obviously the only contribution.

For the single-step mutation model in particular, we can define *deleterious* and *advantageous mutational effects* separately as $s_k^+ = R_k - R_{k+1}$ and $s_k^- = R_{k-1} - R_k$, respectively. For decreasing fitness values (which is the usual case, but not strictly presupposed here) these are positive. This way we obtain

$$V_R = \overline{s^+ U^+ - s^- U^-} = \overline{s^+}\;\overline{U^+} - \overline{s^-}\;\overline{U^-} + \mathrm{Cov}(s^+, U^+) - \mathrm{Cov}(s^-, U^-) \tag{20}$$

for the equilibrium variance, a result we will rely on in the following.

Analogously, we define the population mean, $\bar{X}(t) = \sum_{i=0}^{N} X_i\, p_i(t)$, and variance, $V_X(t) = \sum_i (X_i - \bar{X}(t))^2\, p_i(t)$, of the mutational distance.

We will also need the average of our observables with respect to the ancestral distribution defined in (10), $\hat{O}(\tau, t) := \sum_i O_i\, a_i(\tau, t) = \sum_i z_i(\tau, t)\, O_i\, p_i(t)$, the *ancestral average*. In the following, we will only be concerned with the ancestral distribution in equilibrium, i.e., with both t and τ going to infinity. For irreducible $\boldsymbol{M}$, this is given by

$$\hat{O} := \sum_i O_i\, a_i = \sum_i z_i\, O_i\, p_i\,. \tag{21}$$

These averages may be read forward in time (corresponding to a weighting of the current population with expected offspring numbers), and backward in time (corresponding to an averaging with respect to the distribution of the ancestors). A third interpretation is available if $\boldsymbol{M}$ is irreducible, which entails that the equilibrium backward process defined by $\boldsymbol{Q}$ is ergodic (see the end of Section 2.2). Then, with probability 1, the equilibrium ancestral average also coincides with the average of the observable over a lineage backwards in time by Birkhoff's ergodic theorem.

Note that the information so obtained is not contained in the population average, which is merely a 'time-slice' average. The ancestral mean adds a time component to the averaging procedure, which provides extra information on the evolutionary dynamics. In [Her02, App. A] it is shown that the ancestral averaging coincides with the way observables are evaluated in a system of quantum statistical mechanics.

2.5 Linear response and mutational loss

We now come to another interpretation of the equilibrium ancestral frequencies introduced in Section 2.2. Consider the derivative of the equilibrium mean fitness with respect to the i-th fitness value in a general system of parallel mutation and selection (2),

$$\frac{\partial \bar{R}}{\partial R_i} = \frac{\partial}{\partial R_i}\Big(\sum_{j,k} z_j H_{jk} p_k\Big) = a_i + \bar{R}\frac{\partial}{\partial R_i}\Big(\sum_j z_j p_j\Big) = a_i\,. \tag{22}$$

Here, we made use of the normalization condition $\sum_j z_j p_j = \sum_j a_j \equiv 1$. The ancestral frequency a_i therefore measures the *linear response* (or *sensitivity*) of the equilibrium mean fitness to changes in the i-th fitness value.[9] A similar calculation for the response to changes in the mutation rates results in

$$\frac{\partial \bar{R}}{\partial M_{ij}} = (z_i - z_j)\, p_j\,. \tag{23}$$

Using (22) and (23), we can express the equilibrium mean fitness as

$$\bar{R} = \hat{R} + \sum_{i,j} z_i M_{ij} p_j = \sum_i R_i \frac{\partial \bar{R}}{\partial R_i} + \sum_{i,j} M_{ij}\frac{\partial \bar{R}}{\partial M_{ij}}\,. \tag{24}$$

[9]If mutation is coupled to reproduction, the linear response to variations in the death rate D_i is given by $-a_i$.

Let us give a variational interpretation for the ancestral mean fitness as well. To this end, we define a quantity G as the difference of ancestral and population mean fitness in equilibrium. Assume now that we change all mutation rates M_{ij} by variations in a common factor μ. From (24) and (22) we then find that the mutational loss relates to the linear response of the equilibrium mean fitness to changes in the mutation rates as

$$G := \hat{R} - \bar{R} = -\mu \frac{\partial \bar{R}}{\partial \mu} \,. \tag{25}$$

Actually, this relation holds for arbitrary (haploid) mutation–selection systems, in particular also if mutation and reproduction are coupled (in which case the mutation *rates* are replaced by mutation *probabilities*).

There is a second line of interpretation, which clarifies the role of G in the equilibrium dynamics. If an individual mutates from j to i, its offspring expectation changes by $z_j - z_i$, where the sign determines whether a loss (+) or gain (−) is implied. Since the mutational flow from j to i in equilibrium is $M_{ij}\, p_j$, the entire system loses offspring at rate $\sum_{i,j}(z_j - z_i)\, M_{ij}\, p_j$, which is the same as G (compare with (24)). Hence, we refer to G as the *mutational loss* of the system.

The mutational loss does not include any information about the destination of the 'lost' offspring. This, however, may easily be found by recalling that, asymptotically, every ancestor of type i leaves a fraction of $z_i p_j$ descendants of type j in the equilibrium population. Furthermore, $p_i(z_i - 1) = a_i - p_i$ is the excess offspring produced by an i-individual. We thus come to a picture of a constant flow of mutants from the ancestors to the equilibrium population.

2.6 Three limiting cases

For many of the results and all examples, we will restrict our treatment to the case of the single-step mutation model as described by (3). Although most results do not depend on this particular choice, we will, for simplicity, concentrate on this scheme here, and only briefly mention possible extensions. Some discussion of this model with respect to its approximation of 'real' biological systems is given in [Her02, Sec. 2.6].

Our primary aim in the following section is to establish simple relations for the equilibrium means and variances of mutational distance and fitness. Whereas these relations are, in general, approximations, they hold as exact identities in three limiting cases. All three are biologically meaningful by themselves, and two of them are indeed well studied.

For a consistent treatment, it will be advantageous to think of the fitness values and mutation rates as being determined by the mutational distance *per class* (or site), $x_k := X_k/N = k/N \in [0, 1]$,

$$R_k = N r_k = N r(x_k) \,, \qquad U_k^{\pm} = N u_k^{\pm} = N u^{\pm}(x_k) \,. \tag{26}$$

Here, also r_k and $u_k^{\pm}$ are introduced as fitness and total mutation rates per class. They can now be thought of as being defined, without loss of generality, by three functions r and $u^{\pm}$ on the compact interval $[0, 1]$. We will refer to r as the *fitness function*, and to u^+ and u^- as the (deleterious and advantageous) *mutation functions* of the model.

Both u^+ and u^- are assumed to be continuous and positive, with boundary conditions $u^-(0) = u^+(1) = 0$, and r to be bounded from above and to have at most finitely many discontinuities, being either left or right continuous at each discontinuity in $]0,1[$. Furthermore, at each point x of discontinuity of r, we assume $r(x)$ to be the larger of the left- and right-sided limit values, respectively $r(0) \geq r(0^+)$ and $r(1) \geq r(1^-)$ at the interval boundaries. (Here, $-\infty$ is explicitly allowed for the lower value.) This should include all biologically relevant examples. For the biallelic model, the mutation functions are simple linear functions of x,

$$u^+(x) = \mu(1+\kappa)(1-x)\,, \qquad u^-(x) = \mu(1-\kappa)x\,. \tag{27}$$

Note that the classical stepwise mutation model [Oht73] is not covered by this framework, since its genotype space $\mathbb{Z}$ is inherently non-compact. However, if it has a proper (i.e., non-zero) equilibrium genotype distribution, it may be approximated by a model with finitely many genotypes, and thus, indirectly, the above procedure applies.

The first exact limiting case is given by *unidirectional mutation*, defined as $u^- \equiv 0$ in our model. The second one is the *linear case*, in which fitness and mutation rates depend linearly on some trait $Y_k = Ny_k = Ny(x_k)$ with $y(0) = 0$ and $y(1) = 1$, such as

$$r(x) = r_0 - \alpha y(x)\,, \qquad u^+(x) = \beta^+(1 - y(x))\,, \qquad u^-(x) = \beta^- y(x)\,, \tag{28}$$

with strictly positive constants $\beta^\pm$. Note that, if $y(x)$ is equal to the mutational distance x, the fitness function is linear and the mutation functions $u^\pm$ reproduce the mutation scheme of the biallelic model if $\beta^\pm = (1 \pm \kappa)\mu$. This case can be understood as the limit of vanishing epistasis, in which the system is known as the Fujiyama model in the sequence space literature, cf. [Kau93].

The third case is the limit of an infinite number of mutation classes, $N \to \infty$, which we will call *mutation class limit* for short. In the case of the biallelic multilocus model, this limit has been used and discussed in [Baa01]. Biologically, it addresses the situation of weak or almost neutral mutations, where the average mutational effect (over the mutation classes) is small compared to the mean total mutation rate, $U \gg s$. The limit further assumes that differences in mutation rate between neighboring (pairs of) classes are small compared to the mean rate itself. In this case, genetic change by mutation proceeds in many steps of small average effect and the model is a genuine multi-class model in the sense that typically a large number of classes are relevant in mutation–selection equilibrium. Note that only the *average* mutational effect must be small; this allows for single steps with much larger effect (such as in truncation selection, see Figure 10).

Technically, the limit $N \to \infty$ is performed such that the mutational effects $s^\pm$ and the fitness values and mutation rates *per class*, r and $u^\pm$, remain constant. If fitness values and mutation rates are defined by the three functions r and $u^\pm$ as described above (26), increasing N simply leads to finer 'sampling' of these functions.

With this kind of scaling, the means and variances *per class* of the observables defined in Section 2.4 approach well defined limits (cf. Section 3), which then serve as approximations for the original model with finite N. We will denote them by the corresponding lower case letters, i.e., $\hat{r} := \hat{R}/N$, $v_X := V_X/N$, etc.; an additional subscript will indicate the limit value, e.g., $\bar{x}_\infty := \lim_{N\to\infty} \bar{x}$. Note that it is, in general, the *variance per class* of a given quantity that is meaningful in this limit, not the variance of the *quantity per class* (e.g., $\mathrm{Var}(X/N)$), which tends to zero (cf. Section 3.6). The described limit is the biological analog of the *thermodynamic limit* in statistical mechanics. It is, however, substantially different from the so-called *infinite-sites limit*, in which the stepwise mutation model is obtained. Both issues are discussed in [Her02, App. A, Sec. 2.7].

Another point worth mentioning is that, for the mutation class limit, no limiting model exists. Although the limiting genotype space is a compact interval, one may not define a sensible continuum-of-alleles (COA) model because, in the limit, the mutant distributions concentrate on the source genotype. Nevertheless, a model with a finite number of genotypes may approximate a COA model (and vice versa), as discussed in the next chapter.

3 Results for means and variances of observables

This section is devoted to our main findings for the single-step mutation model, which are summarized in Section 3.1. The proofs and a more extended discussion are postponed to Sections 3.2–3.6. Section 3.7 then contains some remarks about the accuracy of these results for models not among the exact limiting cases described in Section 2.6.

3.1 Statement of the results

Let us start by recollecting the main definitions and assumptions from Section 2. We think of the system as being defined by a fitness function $r\colon [0,1] \to \mathbb{R}$, mutation functions $u^{\pm}\colon [0,1] \to \mathbb{R}_{\geq 0}$, and the number of mutation classes, N. Here, r is assumed to have at most *finitely* many discontinuities, being either left or right continuous at each discontinuity $x \in \,]0,1[$, with $r(x) = \max\{r(x^-), r(x^+)\}$, and satisfying $r(0) \geq r(0^+)$, $r(1) \geq r(1^-)$.[10] Further, $u^{\pm}$ are taken to be continuous and positive, with boundary conditions $u^-(0) = u^+(1) = 0$. The trait values are defined as $x_k = k/N$ ($0 \leq k \leq N$).

Then, the population frequencies are given by the PF eigenvector $\boldsymbol{p}$ corresponding to the eigenvalue equation

$$\big(r(x_k) - u^+(x_k) - u^-(x_k)\big)\, p_k + u^+(x_{k-1})\, p_{k-1} + u^-(x_{k+1})\, p_{k+1} = \bar{r}\, p_k\,, \qquad (29)$$

with the (population) mean fitness $\bar{r}$ as PF eigenvalue. (Here, the connection to the evolution equation (3) is given via $R_k = N r(x_k)$ and $U_k^{\pm} = N u^{\pm}(x_k)$.) The ancestral frequencies $\boldsymbol{a}$ can be determined from the PF eigenvector of the symmetrized equation

$$\begin{aligned}\big(r(x_k) - u^+(x_k) - u^-(x_k)\big)\sqrt{a_k}&\\ + \sqrt{u^+(x_{k-1})u^-(x_k)}\sqrt{a_{k-1}} + \sqrt{u^+(x_k)u^-(x_{k+1})}\sqrt{a_{k+1}} &= \bar{r}\sqrt{a_k}\,,\end{aligned} \qquad (30)$$

cf. (17). The mean fitness and trait values with respect to both frequencies are defined as

$$\bar{r} = \sum_k r(x_k)\, p_k\,, \quad \bar{x} = \sum_k x_k\, p_k\,, \quad \hat{r} = \sum_k r(x_k)\, a_k\,, \quad \hat{x} = \sum_k x_k\, a_k\,. \qquad (31)$$

[10] Note that a *lower* limit value of $-\infty$ is allowed in each case.

We consider three limiting cases. For unidirectional mutation, $u^- \equiv 0$ is assumed instead of positivity, the linear case is given by (28), and the mutation class limit is defined as $N \to \infty$. In the latter case, we use indices $\bar{r}_N$, $\bar{r}_\infty$ etc. to denote the finite-size and limit values, respectively.

The main result is

Theorem 1 (maximum principle). *Let*

$$g(x) = u^+(x) + u^-(x) - 2\sqrt{u^+(x)u^-(x)}\,. \tag{32}$$

(a) *In the mutation class limit,*

$$\bar{r}_\infty := \lim_{N\to\infty} \bar{r}_N = \sup_{x\in[0,1]} \big(r(x) - g(x)\big)\,. \tag{33}$$

If the supremum is attained at a unique value (under the above assumptions, there is at least one), then this is precisely the ancestral mean $\hat{x}_\infty := \lim_{N\to\infty} \hat{x}_N$ and we have

$$\bar{r}_\infty = r(\hat{x}_\infty) - g(\hat{x}_\infty) = \hat{r}_\infty - g(\hat{x}_\infty)\,. \tag{34}$$

In any case, when increasing N, the ancestral distribution may only concentrate near those values of x for which the supremum is attained.

(b) *In the linear case* (28),

$$\bar{r} = \max_{y\in[0,1]} \big(r(y) - g(y)\big)\,, \tag{35}$$

where, for strictly positive constants $\beta^\pm$ in (28)*, the maximum is attained at a unique value, which equals the ancestral mean trait $\hat{y}$, and $r(\hat{y}) = \hat{r}$ holds.*

(c) *For unidirectional mutation, where $g = u^+$, the equilibrium population with maximal mean fitness is characterized by*

$$\bar{r} = \max_k \big(r(x_k) - u^+(x_k)\big)\,. \tag{36}$$

Here, the largest value of k at which the maximum is attained, $\hat{k}$, defines the only non-zero ancestral frequency $a_{\hat{k}} = 1$, which yields $\hat{x} = x_{\hat{k}}$ and $r(\hat{x}) = \hat{r}$. (However, if the maximum is not unique, the p_k and z_k are mutually singular; hence, in this case, the ancestral frequencies can not be constructed as $a_k = z_k p_k$, which is identically zero.)

The function g, defined as twice the difference between the arithmetic and geometric mean of the mutation functions, will be called *mutational loss function* due to

Proposition 1. *In the mutation class limit, if* (34) *holds, the mutational loss per class, $g_N = G_N/N$, converges to $g(\hat{x}_\infty)$. In the linear case and for unidirectional mutation, we have $G/N = g(\hat{y})$, respectively $G/N = g(\hat{x})$.*

For the biallelic model, the mutational loss function reads explicitly

$$g(x) = \mu\left(1 + \kappa - 2\kappa x - 2\sqrt{(1-\kappa^2)x(1-x)}\right)\,. \tag{37}$$

The population mean of the trait value and the variances are given by

Theorem 2. *For the mutation class limit, assume r to be continuously differentiable with derivative r'. Then, in this limit and the linear case, we have*

$$\bar{r}_\infty = r(\bar{x}_\infty)\,, \qquad \textit{respectively} \qquad \bar{r} = r(\bar{y})\,, \tag{38}$$

if this equation has a unique solution (e.g., for strictly monotonic r, i.e., $\alpha \neq 0$ in the linear case). If further, in the linear case, $y(x) = x$, the variances per site of fitness and of distance from the wildtype are given by

$$v_{R,\infty} = -r'(\bar{x}_\infty)\big(u^+(\bar{x}_\infty) - u^-(\bar{x}_\infty)\big) \qquad \textit{and} \qquad v_{X,\infty} = \frac{v_{R,\infty}}{\big(r'(\bar{x}_\infty)\big)^2}\,, \tag{39}$$

respectively, without the index ∞ for the linear case. Here, $-r'(\bar{x}_\infty)$ is (the limit of) the population mean of the mutational effects.

If r has a jump discontinuity at x_{jump} from r^+ to r^- and we have $r^+ \leq \bar{r}_\infty \leq r^-$, then $\bar{x}_\infty = x_{\text{jump}}$ and $v_{R,\infty}$ diverges. In this case, $V_{r,\infty} = \lim_{N\to\infty} V_R/N^2$ is finite,

$$V_{r,\infty} = (r^+ - \bar{r}_\infty)(\bar{r}_\infty - r^-)\,. \tag{40}$$

For the biallelic model, (39) reads explicitly

$$v_{R,\infty} = -r'(\bar{x}_\infty)\mu\,(1 + \kappa - 2\bar{x}_\infty) \qquad \text{and} \qquad v_{X,\infty} = -\frac{\mu\,(1 + \kappa - 2\bar{x}_\infty)}{r'(\bar{x}_\infty)}\,. \tag{41}$$

Concerning (40), see the examples in Figures 8 and 10.

Note that, if the position x_{opt} of the fitness optimum lies in the interior of $[0, 1]$ (i.e., stabilizing rather than directional selection is considered) and if mutation is symmetric between adjacent classes[11] around it (which is often assumed for stabilizing selection), i.e., $u^+(x) = u^-(x)$ for all x in some neighborhood of x_{opt}, we have $g(x) = 0$ there. Then, in the mutation class limit, the above results trivially yield $\hat{x}_\infty = \bar{x}_\infty = x_{\text{opt}}$ and $v_{R,\infty} = v_{X,\infty} = 0$. The latter implies that the next weaker order of the variances, $V_{R,\infty}$, respectively $V_{X,\infty}$, is relevant, about which our results provide no information.

The results presented here lead to simple graphical constructions of the means as shown in Figure 4. These allow for an intuitive overview over the dependence of these quantities on (the shape of) the fitness and mutation functions, without the need for explicit calculations.

We now come to the proofs and some interpretation. Our starting point is the mutation–selection equilibrium of the single-step mutation model (3) for finite N, i.e., the eigenvalue equation (29). We will mostly use the equivalent equation (30) for the ancestral distribution, which is the eigenvalue equation for the largest eigenvalue of the symmetric matrix $\tilde{\boldsymbol{H}}$ (divided by N). For the latter, Rayleigh's principle is applicable, which is a general maximum principle involving the full $(N+1)$-dimensional space: $\bar{r} = N^{-1} \sup_{\boldsymbol{y} \neq 0} \sum_{k,\ell} y_k \tilde{H}_{k\ell} y_\ell / \sum_k y_k^2$. In Sections 3.2–3.4 we will show, for each of the three limiting cases separately, how it boils down to the simple scalar maximum principle of Theorem 1. We will then come to the proofs of Proposition 1 and Theorem 2 in Sections 3.5 and 3.6, respectively.

[11] This is not to be confused with symmetric site mutation in the biallelic model, described by $\kappa = 0$.

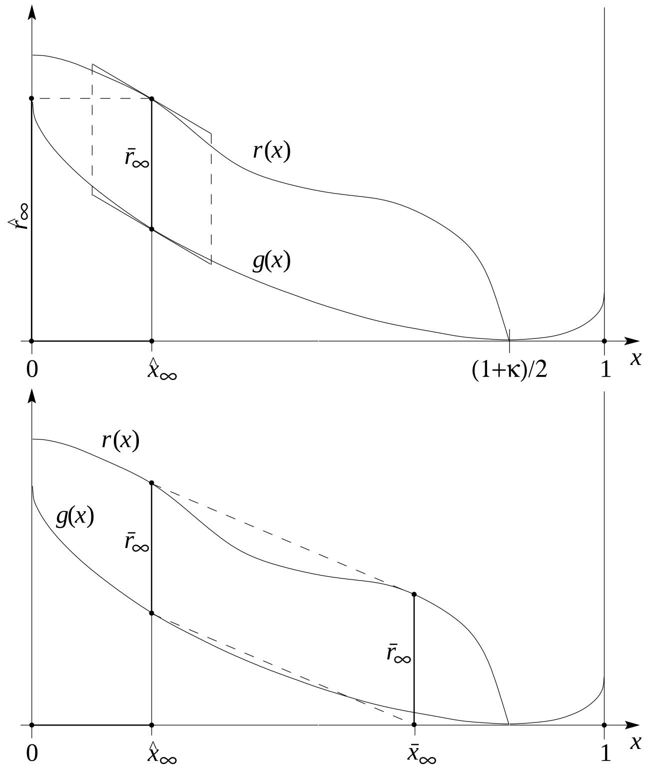

Figure 4: Graphical constructions for the observable means in the mutation class limit, following the results in Section 3.1. Upper part: $\bar{r}_\infty$ is the maximal distance $r(x) - g(x)$, cf. (33). This is attained at $x = \hat{x}_\infty$, cf. (34), where $r'(\hat{x}_\infty) = g'(\hat{x}_\infty)$. Lower part: $\bar{x}_\infty$ is the solution of $\bar{r}_\infty = r(\bar{x}_\infty)$, cf. (38).

3.2 Unidirectional mutation

We start with the limiting case of unidirectional mutation, since exclusion of back mutations leads to a considerably simpler situation, and we can show how our findings connect to well-known results. To be specific, we assume

$$u^- \equiv 0 \qquad \text{and} \qquad u^+(x) > 0 \quad \text{for } x \in [0, 1[. \tag{42}$$

All results then follow fairly directly from the equilibrium condition (29).

Owing to $u^- \equiv 0$ and the resulting reducibility of $\boldsymbol{H}$, the equilibrium distribution $\boldsymbol{p}$ is in general not unique (compare [Sch74, Sec. I.2]). We therefore require $\bar{r}$ to be the largest eigenvalue of (29). The following lemma then ensures the uniqueness of $\boldsymbol{p}$, which is always attained if the initial population satisfies $p_0(0) > 0$ (see [Wil65, Ch. 9]).

Lemma 1. *For any non-negative eigenvector $\boldsymbol{p}$ of (29) with $\|\boldsymbol{p}\|_1 = 1$ and eigenvalue $\bar{r}$, there exists a label $\hat{k}$, $0 \le \hat{k} \le N$, which divides all classes of genotypes into two parts,*

$$p_k = 0 \quad \textit{for } k < \hat{k}, \qquad p_k > 0 \quad \textit{for } k \ge \hat{k}\,, \tag{43}$$

and satisfies

$$r(x_{\hat{k}}) - u^+(x_{\hat{k}}) = \bar{r}\,. \tag{44}$$

Furthermore,

$$r(x_k) - u^+(x_k) < \bar{r} \quad \textit{for } k > \hat{k}, \tag{45}$$

which makes $\boldsymbol{p}$ *the only eigenvector to* $\bar{r}$ *with the above properties. In other words, if condition* (44) *is true for more than one label, then* $\hat{k}$ *is the largest of them.*

Proof: The first statement follows directly from (29) and (42): If $\hat{k}$ denotes the smallest label such that $p_{\hat{k}} > 0$, then $p_k > 0$ for all $k > \hat{k}$. For the second statement, assume there is a label $k > \hat{k}$ with $r(x_k) - u^+(x_k) \geq \bar{r}$. Then (29) and (43) lead to the contradiction $\bar{r} = r(x_k) - u^+(x_k) + u^+(x_{k-1})p_{k-1}/p_k > \bar{r}$. □

A similar result is obtained for the corresponding left eigenvector $\boldsymbol{z}$ and a label $\check{k}$,

$$z_k = 0 \quad \text{for } k > \check{k}, \qquad z_k > 0 \quad \text{for } k \leq \check{k}\,. \tag{46}$$

However, if (43) is satisfied for more than one label, then $\check{k}$ is the smallest such label, and thus $\check{k} < \hat{k}$. As a consequence, we have $z_k = 0$ whenever $p_k > 0$ (and vice versa). Thus, the interpretation of the z_k as the expected relative numbers of offspring is invalid and the ancestral frequencies can not be constructed as $a_k = z_k p_k$ (which is identically zero). But, in any case, the mutational distance of every line of ancestors in equilibrium dynamics converges to $\hat{k}$ (with probability 1). Thus, the only non-zero element of the ancestral distribution is $a_{\hat{k}} = 1$.

With these results, we are able to give the

Proof of Theorem 1(c): Let $\bar{r}$ be chosen according to (36), $\hat{k}$ as the largest k at which the maximum is attained, $p_k = 0$ for $k < \hat{k}$, and $p_{\hat{k}} = C > 0$. Then, (29) uniquely defines all p_k for $k > \hat{k}$, and we may choose C such that $\|\boldsymbol{p}\|_1 = 1$. This way, we constructed an eigenvector $\boldsymbol{p}$ for the eigenvalue $\bar{r}$. According to Lemma 1, this must be the largest eigenvalue and $\boldsymbol{p}$ is unique. The other statements have been discussed above. □

If the sequence $r(x_k)$ *or* the sequence $u^+(x_k)$ is monotonically decreasing (as in the biallelic model), $\hat{k}$ is also the fittest class present in the equilibrium population,

$$\hat{r} = r(\hat{x}) = \max_k \{r(x_k) : p_k \neq 0\}\,. \tag{47}$$

If additionally $\hat{k}$ coincides with the class of maximal fitness, i.e., $\hat{r} = r_{\max}$, then (44) is a special case of Haldane's principle, which relates the mutation load l to the deleterious mutation rate of the fittest class [Kim66, Bür98],

$$l = r_{\max} - \bar{r} = u^+(\hat{x})\,. \tag{48}$$

In derivations of (variants of) this equation, it is often tacitly assumed that the equilibrium frequency of the fittest class is non-zero. This, however, is in general not the case and must be made explicit here since we are also interested in the change of the equilibrium distribution with varying mutation rates. This can lead to a shift in $\hat{k}$ and hence in $\hat{r}$.

3.3 The linear case

If fitness values and mutation rates depend linearly on some trait Y, as described in (28), the maximum principle holds as an exact identity. This may be derived from (30) by a short direct calculation.

Proof of Theorem 1(b): We show that the system (30) reduces to just two equations, one corresponding to the necessary extremum condition following from (35), the other establishing that $\bar{r}$ indeed equals this maximum.

Taking the difference of two arbitrary equations of the linear system (30), say for k and ℓ, divided by $\sqrt{a_k}$ and $\sqrt{a_\ell}$, respectively, we get

$$
\begin{aligned}
&(\beta^+ - \beta^- - \alpha)(y_k - y_\ell) \\
&\qquad + \sqrt{\beta^+\beta^-}\Bigg(\sqrt{y_k(1-y_{k-1})}\sqrt{\frac{a_{k-1}}{a_k}} - \sqrt{y_\ell(1-y_{\ell-1})}\sqrt{\frac{a_{\ell-1}}{a_\ell}} + \\
&\qquad\qquad \sqrt{y_{k+1}(1-y_k)}\sqrt{\frac{a_{k+1}}{a_k}} - \sqrt{y_{\ell+1}(1-y_\ell)}\sqrt{\frac{a_{\ell+1}}{a_\ell}}\Bigg) = 0\,.
\end{aligned} \tag{49}
$$

With the ansatz

$$
\frac{a_{k-1}}{a_k} = C\frac{y_k}{1-y_{k-1}} \quad\Leftrightarrow\quad \frac{a_{k+1}}{a_k} = C^{-1}\frac{1-y_k}{y_{k+1}}\,, \tag{50}
$$

Equation (49) can be divided by $(y_k - y_\ell)$ and becomes independent of k and ℓ. Note that (50) also takes care of the boundary conditions $a_{-1} = a_{N+1} = 0$ if $y_0 = 0$ and $y_N = 1$. Summing both sides of $(1-y_{k-1})a_{k-1} = Cy_k a_k$ over k, we obtain $C = (1-\hat{y})/\hat{y}$ and thus from (49)

$$
\beta^+ - \beta^- - \alpha + \sqrt{\beta^+\beta^-}\frac{1-2\hat{y}}{\sqrt{\hat{y}(1-\hat{y})}} = 0\,, \tag{51}
$$

which is exactly the extremum condition $r'(\hat{y}) = g'(\hat{y})$ following from (33). Together with the negative second derivative, this implies the maximum principle. It is then straightforward to show that the solution of (51) is unique. As a consequence, the maximum in (35) is indeed assumed at the ancestral mean trait value $\hat{y}$.

Further, we can use (50) to eliminate $a_{k\pm1}$ from (30). After multiplication by $\sqrt{a_k}$ this reads

$$
\left[r_0 - \alpha y_k - \bar{r} - \beta^+(1-y_k) - \beta^- y_k + \sqrt{\beta^+\beta^-}\left(y_k\sqrt{\frac{1-\hat{y}}{\hat{y}}} + (1-y_k)\sqrt{\frac{\hat{y}}{1-\hat{y}}}\right)\right]a_k = 0 \tag{52}
$$

and we obtain, by summation over k,

$$
\bar{r} = r_0 - \alpha\hat{y} - \beta^+(1-\hat{y}) - \beta^-\hat{y} + 2\sqrt{\beta^+\beta^-\hat{y}(1-\hat{y})} \;=\; r(\hat{y}) - g(\hat{y})\,, \tag{53}
$$

so the mean fitness is indeed given by (35). Since fitness is assumed linear in the trait, the mean values with respect to the population and ancestral distributions are also related via $\bar{r} = r(\bar{y})$ and $\hat{r} = r(\hat{y})$. □

For an interpretation of this result, first consider a trait proportional to the mutational distance from the reference class, in which case the system coincides with the Fujiyama model. Since this is a model without epistasis, the means and variances are easily obtained [O'B85, Baa01]. In particular, they are independent of the number of classes. What is more, our derivation shows that they only rely on a linear dependence of fitness and mutation functions on some trait, as well as the boundary conditions for the mutation functions. This means that they remain unchanged if mutation classes are permuted, or even subjoined or removed.

3.4 Mutation class limit

The main idea of the proof of the maximum principle in the limit $N \to \infty$ is to look at the system *locally*, i.e., at some interval of mutation classes in (29) and (30). This will provide us with upper and lower bounds for the mean fitness of a system with finite N. In the limit $N \to \infty$, these can then be shown to converge to the same value $\bar{r}_\infty$.

Proof of Theorem 1(a): For a lower bound, we consider submatrices of $\boldsymbol{H}$ that, for any class $\mathcal{X}_k$, consist of the rows (and columns) corresponding to $\mathcal{X}_{k-m}$ through $\mathcal{X}_{k+n}$. Each of them describes the evolution process on a certain interval of mutation classes at whose boundaries there is mutational flow out, but none in. Thus, each largest eigenvalue, $\bar{r}_{k,m,n}$, corresponding to the local growth rate, is a lower bound for $\bar{r}_N$, compare [Sch74, Cor. of Thm. I.6.4]. In order to estimate $\bar{r}_{k,m,n}$, it is advantageous to use the formulation in ancestor form—with the same local growth rates as largest eigenvalues of the corresponding symmetric submatrices of $\tilde{\boldsymbol{H}}$. Here, lower bounds can be found with Rayleigh's principle, and follow from evaluating the corresponding quadratic form for the vector $(1, 1, \dots, 1)^T$:

$$\bar{r}_N \geq \bar{r}_{k,m,n} \geq \frac{1}{n+m+1} \left(\sum_{\ell=k-m}^{k+n} r_\ell - g_{N,\ell} - \sqrt{u^+_{k-m-1} u^-_{k-m}} - \sqrt{u^+_{k+n} u^-_{k+n+1}} \right), \tag{54}$$

where $g_{N,\ell} = u^+_\ell + u^-_\ell - \sqrt{u^+_{\ell-1} u^-_\ell} - \sqrt{u^+_\ell u^-_{\ell+1}}$. The RHS is itself greater than or equal to

$$\rho_{k,m,n} := \inf_{y \in I_{k,m,n}} \big(r(y) - g(y) \big) - \sup_{y \in I_{k,m,n}} \big| g(y) - g_N(y) \big| - \frac{\sqrt{u^+_{k-m-1} u^-_{k-m}} + \sqrt{u^+_{k+n} u^-_{k+n+1}}}{m+n+1}, \tag{55}$$

where $I_{k,m,n} = [\frac{k-m}{N}, \frac{k+n}{N}]$ and the rules for inf/sup have been applied. We will now construct a sequence $\rho_N(x) := \rho_{k_N(x), m_N(x), n_N(x)}$ for each $x \in [0,1]$, using suitable sequences for the indices, such that

$$\lim_{N \to \infty} \rho_N(x) = r(x) - g(x)\,. \tag{56}$$

Equations (54)–(56) will then establish $\liminf_{N \to \infty} \bar{r}_N \geq \sup_{x \in [0,1]} (r(x) - g(x))$.

Note first that, for $x = 0$ or $x = 1$, $\rho_N(x) = \rho_{xN,0,0} = r(x) - g(x)$ holds for arbitrary N. Now, fix $x \in \,]0,1[$. If r is continuous in $[x-d, x]$ for a suitable $d > 0$, let $k_N(x) = \lfloor xN \rfloor$, $m_N(x) = \lfloor d\sqrt{N} \rfloor$, and $n_N(x) \equiv 0$. Otherwise r is continuous in $[x, x+d]$ for some $d > 0$, and we define $k_N(x) = \lceil xN \rceil$, $m_N(x) \equiv 0$, and $n_N(x) = \lfloor d\sqrt{N} \rfloor$. With these choices, the last term in (55) vanishes for $N \to \infty$ since $m_N(x) + n_N(x) \to \infty$, and the enumerator is bounded. So does the supremum term because of the uniform convergence $g_N \to g$: $\sup_{y \in I_{k_N,m_N,n_N}} |g(x) - g_N(x)| \le \sup_{y \in [0,1]} |g(x) - g_N(x)| \to 0$. The latter follows from the uniform continuity of $\sqrt{u^\pm}$ since, in

$$\begin{aligned} &|g(x) - g_N(x)| \\ &= \left| \left(\sqrt{u^+(x - \tfrac{1}{N})} - \sqrt{u^+(x)} \right) \sqrt{u^-(x)} + \sqrt{u^+(x)} \left(\sqrt{u^-(x + \tfrac{1}{N})} - \sqrt{u^-(x)} \right) \right| , \end{aligned} \tag{57}$$

the terms in parentheses vanish uniformly in x as $N \to \infty$ and $\sqrt{u^\pm(x)}$ is bounded. The infimum term in (55), and thus $\rho_N(x)$, converges to $r(x) - g(x)$ since $x_{k_N(x)} \to x$, the function r is continuous in all $I_N \ni x$, and $|I_N| = (m_N(x) + n_N(x))/N \to 0$. This finishes the proof for the lower bound.

For an upper bound, consider a local maximum of the ancestral distribution, i.e., a k^+ such that $a_{k^+} \ge a_{k^+ \pm 1}$ (with the convention $a_{N+1} = a_{-1} = 0$ such a maximum always exists). Evaluating (30) for this k^+ then yields the inequality

$$\bar{r}_N \le r_{k^+} - g_{N,k^+} \le \sup_k \left(r_k - g_{N,k} \right) . \tag{58}$$

In the limit, this establishes $\limsup_{N \to \infty} \bar{r}_N \le \sup_{x \in [0,1]} (r(x) - g(x))$, from which, together with the lower bound from above, the maximum principle (33) follows (including convergence of the sequence $\bar{r}_N$).

We now prove that the ancestral distribution is concentrated around those x for which $r(x) - g(x)$ is maximal, from which (34) follows if the maximum is unique. Multiplying the equilibrium equation in ancestor form (30) by $\sqrt{a_k}$ and summing over k, we get

$$\begin{aligned} \bar{r}_N = \sum_{k=0}^{N} \Big[&\big(r(x_k) - u^+(x_k) - u^-(x_k) \big) a_k \\ &+ \sqrt{u^+(x_{k-1}) u^-(x_k)} \sqrt{a_k a_{k-1}} + \sqrt{u^+(x_k) u^-(x_{k+1})} \sqrt{a_{k+1} a_k} \Big] . \end{aligned} \tag{59}$$

Using $\sqrt{a_k a_{k \pm 1}} \le \frac{1}{2}(a_k + a_{k \pm 1})$, we obtain

$$\bar{r}_N \le \sum_{k=0}^{N} \left(r(x_k) - g_N(x_k) \right) a_k = \hat{r}_N - \widehat{(g_N)}_N , \tag{60}$$

with g_N as defined above. Since $\bar{r}_N \to \bar{r}_\infty$ and $g_N(x) \to g(x)$ uniformly in $x \in [0,1]$, we can find, for any given $\varepsilon > 0$, an N_ε, such that, for all $N > N_\varepsilon$,

$$\sum_{k=0}^{N} \big(r(x_k) - g(x_k) \big) a_k > \bar{r}_\infty - \varepsilon^2 . \tag{61}$$

We now divide this sum into two parts, $\sum_k := \sum_{k_>} + \sum_{k_\le}$. The first part, $\sum_{k_>}$, collects all k with $r(x_k) - g(x_k) > \bar{r}_\infty - \varepsilon$, the second part contains the rest. We then obtain

$$\bar{r}_\infty - \varepsilon^2 < \sum_{k=0}^{N} \big(r(x_k) - g(x_k) \big) a_k \le \bar{r}_\infty \sum_{k_>} a_k + (\bar{r}_\infty - \varepsilon) \sum_{k_\le} a_k = \bar{r}_\infty - \varepsilon \sum_{k_\le} a_k \tag{62}$$

and thus $\sum_{k_\leq} a_k < \varepsilon$. We conclude that, for N sufficiently large, the ancestral distribution is concentrated in those mutation classes for which $r(x) - g(x)$ is arbitrarily close to its maximum, $\bar{r}_\infty$. □

3.5 Mutational loss

Let us now turn to the connection between the mutational loss G and the mutational loss function $g(x)$.

Proof of Proposition 1: The claim follows, in each of the three cases, from (24), (25), and Theorem 1. This is obvious for the linear case and unidirectional mutation. For the mutation class limit, (34) has to be assumed, then $\hat{r}_\infty = r(\hat{x}_\infty)$ holds. □

Recall further that, in the proof of the maximum principle in the mutation class limit in Section 3.4, we obtained $r(x) - g(x)$ as the largest eigenvalue of a local open subsystem around x. If $\bar{r}_\infty$ is the death rate due to population regulation in the entire system, $r(x) - \bar{r}_\infty - g(x)$ is the net growth rate of the subsystem at x. Hence, $g(x)$ must describe the rate of mutational loss due to the flow out of the local system. This can be made more precise within the framework of large deviation theory, which will be presented in a future publication [Baa].

3.6 Mean mutational distance and the variances

Here, we derive and discuss the results for the mean mutational distance and the variances, which hold in the linear case and for $N \to \infty$.

Proof of Theorem 2: If fitness is linear in an arbitrary trait $y(x)$, we immediately have $\bar{r} = r(\bar{y})$. For the variance formulas, we must additionally assume that fitness is linear in the mutational distance, $r(x) = r_{\max} - \alpha x$. Thus, the covariances in the general formula (20) vanish, and $v_R = \alpha\,(\overline{u^+} - \overline{u^-})$. Due to the linearity, this also determines the variance in mutational distance as $v_X = (\overline{u^+} - \overline{u^-})/\alpha$. These relations do not require that $u^\pm(x)$ are linear in x; they reduce to (39) if this is the case.

In the mutation class limit, we assume that r is continuously differentiable and express $v_{R,\infty}$ as the limit variance for increasing system size N, using (20) for $v_{R,N}$,

$$v_{R,\infty} = \lim_{N\to\infty} \sum_{k=0}^{N} \Big(\frac{r_k - r_{k+1}}{N^{-1}}\, u_k^+ - \frac{r_{k-1} - r_k}{N^{-1}}\, u_k^- \Big) p_k = -\overline{r'\,(u^+ - u^-)}_\infty\,. \tag{63}$$

Here, we made use of the fact that the mutational effects converge uniformly to the corresponding values of $-r'$, i.e., the negative slope of the fitness function.

Since r' is bounded, (63) in particular shows that $v_{R,\infty}$ is finite, and hence

$$V_{r,\infty} = \lim_{N\to\infty} \Big[\sum_{k=0}^{N} r_k^2 p_k - \Big(\sum_{k=0}^{N} r_k p_k \Big)^2 \Big] = \lim_{N\to\infty} N^{-1} v_{R,N} = 0\,. \tag{64}$$

For increasing N, the distribution of fitness values *per class* therefore concentrates around $\bar{r}_\infty$. Accordingly, the mean mutational distance in the limit satisfies (38) if this equation has a unique solution, including convergence of the sequence $\bar{x}_N$. With this, the mean of the mutational effects, $\overline{s^\pm}_N$ (see Section 2.4), converges to $-r'(\bar{x}_\infty)$. Furthermore, we have $v_{R,\infty} = -r'(\bar{x}_\infty)\big(u^+(\bar{x}_\infty) - u^-(\bar{x}_\infty)\big)$. The variance in x can then be obtained via the linear approximation $r(x) \simeq r(\bar{x}_\infty) + r'(\bar{x}_\infty)(x - \bar{x}_\infty)$ as $v_{X,\infty} = v_{R,\infty}/(r'(\bar{x}_\infty))^2$.

If there is a jump in the fitness function, v_R diverges according to the above relation, but $V_{r,\infty}$ is finite and determined by the fraction of the population below and above the jump, which yields (40). □

For fitness functions with kinks, the proof is analogous, as long as the left- and right-sided limits of r' remain bounded, and the convergence of the mutational effects is uniform.

3.7 Accuracy of the approximation

We now illustrate the accuracy of the analytical expressions for means and variances given in Section 3.1. To pay respect to the invariance of the equilibrium distributions under scaling of both reproduction and mutation rates with the same factor, we introduce γ as an overall constant for the reproduction rates. In an application, it should be chosen to represent roughly the average effect of a single mutation on the reproduction rate in a mutant genotype (with the maximum number of mutations considered) as compared to the wildtype. This does not exclude the possibility that effects of single mutations may be quite large. In the figures, both reproduction and mutation rates are given in units of this constant, i.e., as r/γ, respectively μ/γ.

Figure 5 displays an example of a biallelic model that deviates from all three exact limiting cases described in Section 2.6, and, for comparison, three modifications that are closer to one of the exact limits each. All numerical values, also in the rest of this article and in Figure 3, are virtually exact and, if not noted otherwise, obtained by the power method [Wil65, Ch. 9], also known as von Mises iteration, with the matrix $\boldsymbol{H}$. For continuous fitness functions, the approximate expressions for the observable means agree with the exact ones up to corrections of order N^{-1} (as indicated by numerical comparison, not shown) or of order $(u^-)^2$ [Her02, Sec. 5.2]. For fitness functions with jumps, the error seems to be at most of order $N^{-1/2}$ (cf. Figure 10); for a jump at $x = 0$ such as in the sharply peaked landscape, however, the corrections to $\bar{r}$ appear to be still of order N^{-1} for the biallelic model (cf. Figure 6).

Further examples, exhibiting more conspicuous features, are shown in Section 4. For most of them, one will also find good agreement of numerical and analytical values for the means for sequences of length $N = 100$; for the variances, however, one sometimes needs longer ones, like $N = 1000$. In the biallelic model, we generally find stronger deviations for higher mutation rates, as in this regime back mutations become more and more important, whereas for small mutation rates, deviations are of linear order in μ.

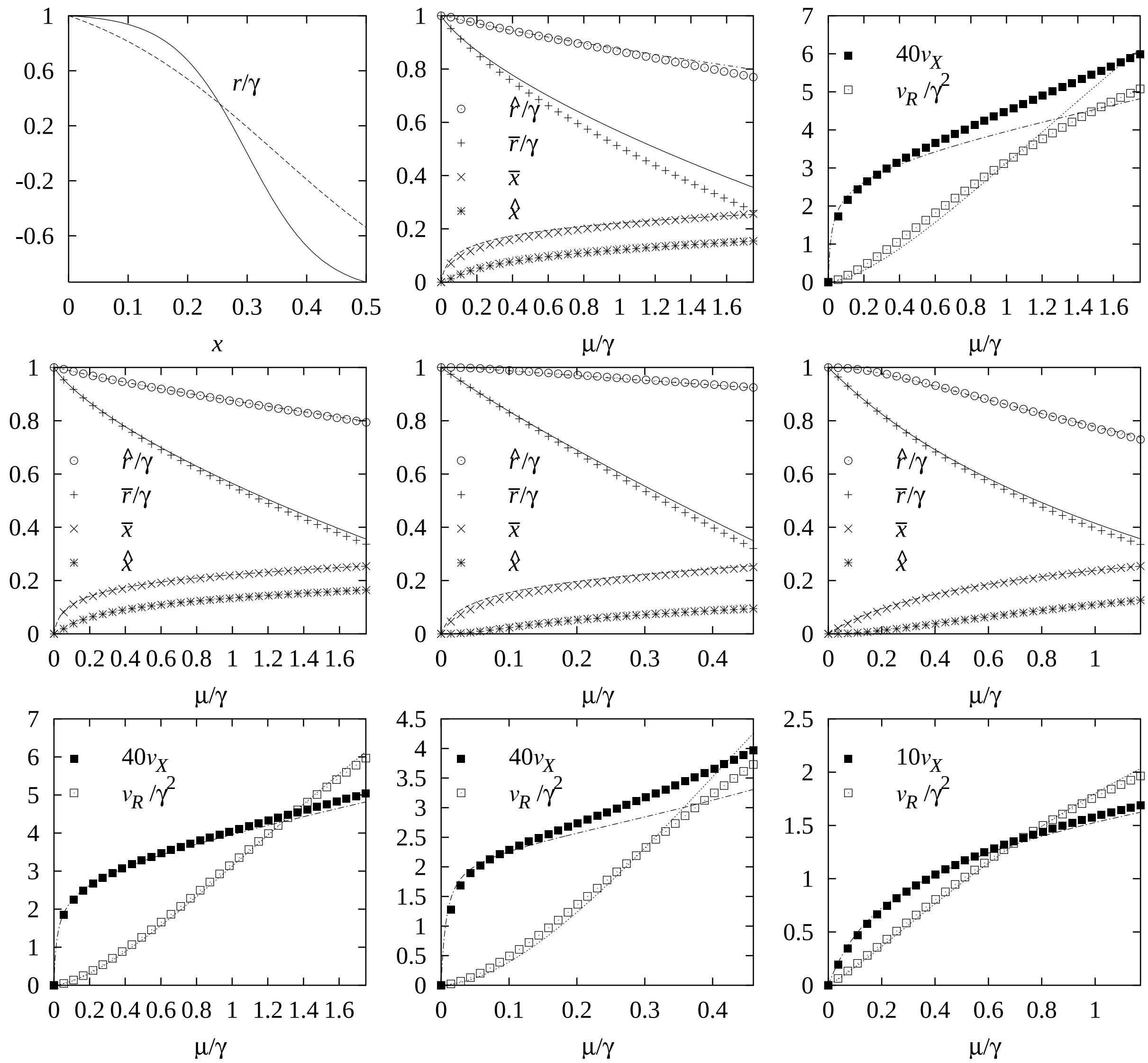

Figure 5: The top row refers to a biallelic model that deviates from all three exact limiting cases described in Section 2.6 in having a strongly non-additive fitness function r/γ (left, solid line), symmetric site mutation ($\kappa = 0$), and small sequence length ($N = 20$). The mean values of the observables (middle) and corresponding variances (right) are shown as a function of the mutation rate μ/γ, both for the model itself (symbols) and according to the expressions given in Section 3.1 (lines, sometimes hidden by symbols). Even here, we find reasonable agreement. Deviations, however, are visible for larger mutation rates. As can be seen from the last two rows, going towards any of the three exact limits, i.e., increasing the number of mutation classes (left, $N = 100$), going to more asymmetric mutation (middle, $\kappa = 0.8$), or using a different fitness function with less curvature (right, r/γ: top left, dashed line), we find that these deviations vanish quickly. In the case of increasingly asymmetric mutation, however, this is not true for the variances, since the approximation becomes only exact here in either of the other two limits (cf. Section 3.6).

4 Application: threshold phenomena

In this section, we analyze how the equilibrium behavior of the single-step mutation model changes if the mutation rates are varied relative to the fitness values. Usually, if mutation rates change slightly, the observable means and variances (e.g., of traits and fitness) at the new equilibrium are close to the old ones. At certain *critical mutation rates*, however, threshold phenomena may occur, associated with much larger effects. The prototype of this kind of behavior is the so-called *error threshold*, first observed in a model of prebiotic evolution many years ago [Eig71] and discussed in numerous variants ever since (for review, see [Eig89, Baa00]).

Here, we will discuss and classify related behavior in our model class. We shall, however, avoid the term error threshold as the collective name for all threshold effects, but rather, and more generally, speak of *mutation thresholds*. This is because the definition of the error threshold is closely linked to the model in which it had been observed originally, namely the quasispecies model with the sharply peaked fitness landscape.

Following [Her02, Sec. 6], we will define mutation thresholds by discontinuous changes in the observable means (or their derivatives) as functions of the mutation rates. Since the largest eigenvalue of $\boldsymbol{H}$ (being simple) and its (properly chosen) right and left eigenvectors depend analytically on the fitness values and mutation rates (compare [Kat80, Sec. II.1]), we need to apply the mutation class limit $N \to \infty$.[12] Throughout this section, we will therefore consider the limit values only, and hence omit the index ∞.

In order to keep the overall shapes of the fitness and mutation functions constant, we vary all mutation rates by a common scalar factor $\mu \geq 0$, chosen as the mean mutation rate over all classes,

$$\mu = \frac{1}{2} \int_0^1 \big(u^+(x) + u^-(x)\big)\, \mathrm{d}x \,. \tag{65}$$

This is consistent with the definition of μ as the mean point mutation rate for the biallelic model, cf. (27) and Figure 1. By slight abuse of notation, we define the shape of the mutational loss function as $g(1, x) = \mu^{-1} g(x)$ (which does not depend on μ, cf. (32)), and introduce μ as a variable parameter via $g(\mu, x) = \mu\, g(1, x)$.

Then, the population mean fitness, as a function of μ, is given by

$$\bar{r}(\mu) = \sup_{x \in [0,1]} \big(r(x) - g(\mu, x)\big) \,. \tag{66}$$

With the assumption from Section 2.6 that at each point of discontinuity of r the larger of the left- and right-sided limit values is attained, there is, for every $\mu \geq 0$, at least one value of x that maximizes $r(x) - g(\mu, x)$. Hence we may define

$$\hat{x}(\mu) = \max\{x \in [0, 1] : \bar{r}(\mu) = r(x) - g(\mu, x)\} \,, \tag{67}$$

which, by Theorem 1, coincides with the ancestral mean genotype if the supremum in (66) is unique. With respect to $\bar{x}(\mu)$, Theorem 2 states that $r(\bar{x}(\mu)) = \bar{r}(\mu)$ is satisfied. This,

[12] This parallels the application of the thermodynamic limit in statistical physics for the definition of phase transitions.

however, may be ambiguous and, unfortunately, we do not have any further information about $\bar{x}(\mu)$. Hence, we are left with the non-constructive definition $\bar{x}(\mu) = \lim_{N\to\infty} \bar{x}_N(\mu)$.

We begin by stating two general properties.

Lemma 2. *The mean fitness $\bar{r}$ is Lipschitz continuous as a function of μ.*

PROOF: The inequality $|\bar{r}(\mu) - \bar{r}(\mu')| \leq |\mu - \mu'| \max_{x\in[0,1]} g(1,x)$ follows easily from (66) and the properties of the supremum. $\square$

Lemma 3. *For such $\mu \geq 0$ for which the supremum in* (66) *is unique, $\hat{x}$ is continuous.*

PROOF: Let $\varepsilon > 0$ be given and consider $I =]\hat{x}(\mu) - \varepsilon, \hat{x}(\mu) + \varepsilon[\cap [0,1]$. Recall that r is at least left or right continuous at $\hat{x}(\mu)$, having only finitely many discontinuities, and that g is continuous and thus bounded. Since, further, the supremum is unique,

$$h = r(\hat{x}(\mu)) - g(\mu, \hat{x}(\mu)) - \sup_{x\in[0,1]\setminus I} \big(r(x) - g(\mu,x)\big) > 0\,. \tag{68}$$

Therefore, for any $\mu' \geq 0$ with $|\mu - \mu'| < \frac{1}{2}h/\max_{x\in[0,1]} g(1,x) =: \delta$, we have

$$\begin{aligned}\sup_{x\in I}\big(r(x) - g(\mu',x)\big) &> \sup_{x\in I}\big(r(x) - g(\mu,x)\big) - \tfrac{1}{2}h \\ &> \sup_{x\in[0,1]\setminus I}\big(r(x) - g(\mu,x)\big) + \tfrac{1}{2}h > \sup_{x\in[0,1]\setminus I}\big(r(x) - g(\mu',x)\big)\,,\end{aligned} \tag{69}$$

from which $\hat{x}(\mu') \in I$ and thus the claim follows. $\square$

Let us further make a couple of assumptions that exclude pathological cases and thus simplify the discussion without narrowing the biological applications.

(A1) The fitness function r has a unique maximum attained at $\hat{x}(0) = \bar{x}(0) =: x_{\min}$, referred to as the wildtype position.

(A2) The smallest $x_{\max} \in]x_{\min}, 1[$ with $g(1, x_{\max}) = 0$ (i.e., $u^+(x_{\max}) = u^-(x_{\max})$) satisfies $\lim_{\mu\to\infty} \hat{x}(\mu) = \lim_{\mu\to\infty} \bar{x}(\mu) = x_{\max}$, and thus describes mutation equilibrium.[13]

(A3) We have $\bar{x}(\mu), \hat{x}(\mu) \in [x_{\min}, x_{\max}]$ for all $\mu \geq 0$.

Note that, for the biallelic model, $x_{\max} = (1+\kappa)/2$.

In the following discussion, we will distinguish four types of mutation thresholds, which all fall together in the original error threshold of the sharply peaked landscape (cf. Figure 6). These are verbally (and somewhat vaguely) described as

1. a kink in the population mean fitness,
2. the loss of the wildtype from the population,
3. complete mutational degradation, and
4. a jump in the population mean of the mutational distance (or some additive trait).

Here, we will restrict ourselves to their mathematical properties. A detailed biological discussion is given in [Her02, Sec. 6].

[13] Note that $g(1, x_{\min}) > 0$ follows from the latter condition together with (A1).

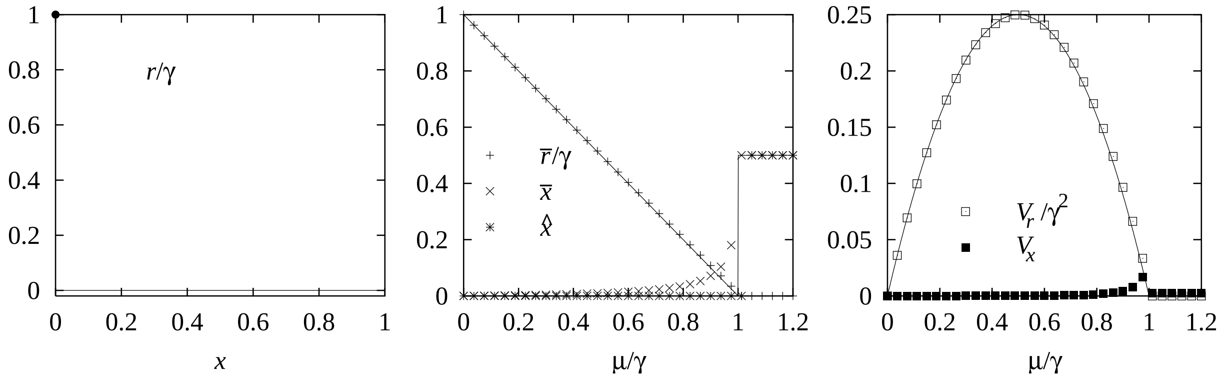

Figure 6: The error threshold of the **sharply peaked landscape** (left) with $r(0) = \gamma$ (bullet) and $r(x) = 0$ for $x > 0$ (line), for the biallelic model with symmetric mutation ($\kappa = 0$). The observable means are shown in the middle, the variances on the right. Symbols correspond to $N = 100$, lines to the expressions in Section 3.1. The ancestral fitness $\hat{r}(\mu)$ (not shown) jumps from γ to 0 at $\mu = \gamma$. Note that V_r follows the scaling described by (63) and is given by (40) for $N \to \infty$.

4.1 Fitness thresholds

It turns out that the kink in the population mean fitness is, in many respects, the most fundamental aspect to classify mutation thresholds. We define a kink in $\bar{r}$ as a non-differentiable point μ_c. Assume $\bar{r}$ to be differentiable in two small enough intervals left and right of such a point μ_c (which is the generic case). There, we conclude from (25) and Proposition 1 that $d/d\mu\, \bar{r}(\mu) = -g(1, \hat{x}(\mu))$. Since $g(1, .)$ is continuous and bounded, the left- and right-sided limits of $\bar{r}'$ at μ_c differ, if they exist, and $\hat{x}$ is discontinuous. According to this reasoning we give the

Definition 1. *A mutation rate μ_c is said to be a* mutation threshold in fitness, *or* fitness threshold *for short, if the mean ancestral genotype $\hat{x}$ is discontinuous at $\mu = \mu_c$.*

As a consequence, the mean mutational distance $\bar{x}$, and the variances v_R and v_X, will typically show kinks as well. An example is presented in Figure 7.

The origin of a fitness threshold is easily understood from the maximum principle. For a generic choice of μ, the function $r(x) - g(\mu, x)$ is maximized for a unique $x = \hat{x}(\mu)$. For some fitness and mutation functions, however, there are particular values of μ that lead to multiple solutions and hence to a jump in $\hat{x}$ when μ is changed across this value.

Theorem 3. *Assume r, $u^{\pm} \in C^2([x_{\min}, x_{\max}])$. Then g is twice continuously differentiable with respect to x in $]x_{\min}, x_{\max}]$. Assume further that $g''(1, x) \neq 0$ whenever $g'(1, x) = 0$. (The prime always denotes the derivative with respect to x.) Consider*

$$C = \sup_{x \in [x_{\min}, x_{\max}]} \left(r''(x) - \frac{r'(x) g''(1, x)}{g'(1, x)} \right) , \tag{70}$$

which might take the values $\pm\infty$. Then there is a fitness threshold if $C > 0$ and there is no fitness threshold if $C < 0$.

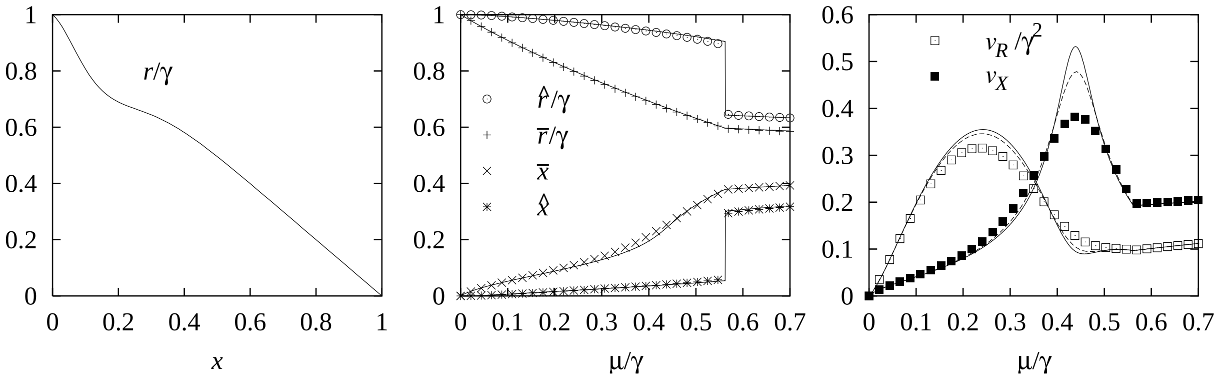

Figure 7: Means (middle) and variances (right) for a biallelic model with asymmetric mutation ($\kappa = 0.4$), and the fitness function r/γ shown on the left. One observes a **fitness threshold** ($\mu_c/\gamma \simeq 0.562$). Symbols correspond to $N = 100$, dashed lines to $N = 500$, and solid lines to the expressions from Section 3.1.

The marginal case $C = 0$ is discussed in [Her02, Sec. 6.2.1], together with extensions to fitness and mutation functions with kinks and jumps. Note that this criterion does not indicate whether there are one or multiple thresholds for a given combination of r and $u^{\pm}$. Neither does it provide the value of $\bar{r}$ at the threshold, nor of μ_c. In fact, the value C is independent of the scalar factor μ, but only depends on the shapes of the mutation and fitness function.

Proof of Theorem 3: The differentiability of g is a direct consequence of the chain rule and the positivity assumptions regarding $u^{\pm}$. With respect to the main statement, recall that, according to Definition 1, a fitness threshold is signaled by a discontinuity in $\hat{x}$. Thus, in the absence of a threshold, $\hat{x}(\mu)$ varies continuously from $x_{\min}$ to $x_{\max}$ by assumptions (A1)–(A3). This is the case if, at each x in the half-open interval $[x_{\min}, x_{\max}[$, the maximum in (66) is uniquely attained for some finite $\mu \geq 0$. It is easily verified that a sufficient condition for this is

$$\forall x \in]x_{\min}, x_{\max}[\;\; \exists \mu > 0 : \quad r'(x) = g'(\mu, x) \quad \text{and} \quad r''(x) < g''(\mu, x)\,. \tag{71}$$

Now assume that $C < 0$, in which case we have, for all $x \in [x_{\min}, x_{\max}]$,

$$r''(x) - \frac{r'(x) g''(1, x)}{g'(1, x)} < 0\,. \tag{72}$$

We first show that both r and g are strictly decreasing in $]x_{\min}, x_{\max}[$. To see this, recall that r has a global maximum at $x_{\min}$ by (A1) and that $g(x) > 0$ for $x \in [x_{\min}, x_{\max}[$ and $g(x_{\max}) = 0$ by (A2). Suppose there exists an $x \in]x_{\min}, x_{\max}[$ with $r'(x) = 0$, and let x_r be the smallest such x. Then either $g'(1, x_r) = 0$ and thus $\lim_{x \nearrow x_r} \Big(r''(x) - \frac{r'(x) g''(1,x)}{g'(1,x)} \Big) = r''(x_r) - r''(x_r) = 0$ by de l'Hospital's rule (which is applicable since, by assumption, $g'' \neq 0$ in a neighborhood of x_r), in contradiction to (72), or $g'(1, x_r) \neq 0$, in which case we obtain $r''(x_r) < 0$ in contradiction to $r'(x) < 0$ for $x \in]x_{\min}, x_r[$. Further, imagine $g'(x) = 0$ for some $x \in]x_{\min}, x_{\max}[$, and let x_g be the largest such x. Then, since $g'(x) < 0$

for $x \in]x_g, x_{\max}[$, we have $g''(x_g) < 0$ (with a strict inequality by assumption) and thus $\lim_{x \searrow x_g} g''(x)/g'(x) = +\infty$, which again contradicts (72) since $r'(x_g) < 0$. Therefore, $\mu(x) := r'(x)/g'(1,x)$ is well-defined and positive everywhere in $]x_{\min}, x_{\max}[$, it guarantees $r'(x) = g'(\mu(x), x)$, and (72) yields $r''(x) < g''(\mu(x), x)$, which completes the first part of the proof by (71).

If $C > 0$ we can find, due to the assumptions made, an x_0 in $]x_{\min}, x_{\max}[$ with $g'(1, x_0) \neq 0$ and $r''(x_0) - r'(x_0)g''(1, x_0)/g'(1, x_0) > 0$. This implies $r''(x_0) - g''(\mu, x_0) > 0$ whenever $r'(x_0) - g'(\mu, x_0) = 0$, i.e., a local minimum. Therefore, the maximum of $r(x) - g(\mu, x)$ is never attained at x_0 and we must have a jump in $\hat{x}(\mu)$. □

4.2 Wildtype thresholds

The loss of the wildtype is the classic criterion for the original error threshold as defined in [Eig71]: For the sharply peaked landscape, the frequency p_0 of the wildtype (or master sequence) remains finite for small mutation rates even for $N \to \infty$, but vanishes above the critical mutation rate (in this limit). The same effect may be observed for any fitness function with a jump at the wildtype position $x_{\min}$.[14]

If r is continuous at $x_{\min}$, however, the population distribution spreads over a large number of mutation classes with similar fitness for any finite mutation rate. While for finite N the frequency in any class remains positive for arbitrary μ (as long as there are back mutations), the frequency of any single mutation class (including the wildtype class) vanishes for $N \to \infty$. Still, one may ask for some related phenomenon that goes together with the loss of the wildtype in all models in which this effect is observed, but defines a threshold also in a broader model class. (The fitness threshold as defined above does not meet this requirement, since fitness functions with a jump at the wildtype may well have multiple fitness thresholds, but only lose their wildtype once.)

Again, we give a definition based on the ancestral distribution.

Definition 2. *A* wildtype threshold *is the largest mutation rate $\mu_c^- > 0$ below which the ancestral mean fitness coincides with the fitness of the wildtype,*

$$\hat{r}(\mu) = \hat{r}(0) = r_{\max}\,, \quad \mu < \mu_c^-\,. \tag{73}$$

This threshold may equivalently be defined as the largest μ_c^- below which $\hat{x}(\mu) = x_{\min}$.

Theorem 4. *There is a wildtype threshold if and only if*

$$\lim_{x \searrow x_{\min}} \frac{g(1,x) - g(1, x_{\min})}{r(x) - r(x_{\min})} < \infty\,. \tag{74}$$

Accordingly, fitness functions with a jump at $x_{\min}$ always lead to a wildtype threshold. Note that the LHS of (74) equals $\lim_{x \searrow x_{\min}} g'(1,x)/r'(x)$ if r and g are differentiable.

[14] As the mean fitness varies continuously, the wildtype frequency in the limit decreases linearly with the mutation rate, until the mean fitness reaches the lower value at the jump. For larger mutation rates, the wildtype frequency in the limit is zero due to the sharpness of the population distribution for $N \to \infty$ (cf. Section 3.6).

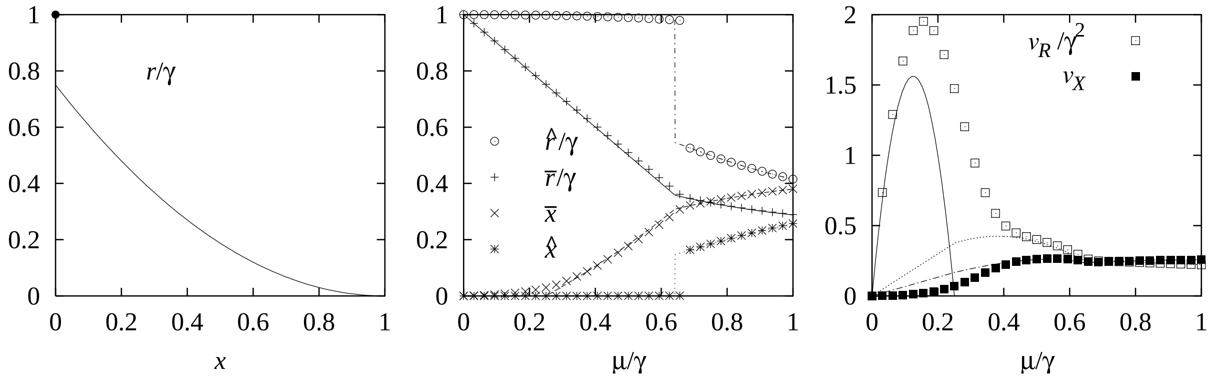

Figure 8: Means (middle) and variances (right) for a model with symmetric mutation ($\kappa = 0$), $N = 100$ (symbols), and the fitness function $r(x) = \frac{3}{4}\gamma\,(1-x)^2$ with an additional single peak of height γ at $x = 0$ (left). Due to the latter, one finds a **wildtype threshold** ($\mu_c^-/\gamma \simeq 0.641$), which is also a fitness threshold. Lines correspond to the expressions in Section 3.1. For $1 \leq \bar{r}/\gamma < \frac{3}{4}$, i.e., $0 \leq \mu/\gamma < \frac{1}{4}$, the variance in fitness no longer follows Equation (41), but scales differently and is given by (40) for $N \to \infty$ (see the discussion in Section 3.6). For finite N, we can approximate v_R by a combination of both relations, where (40) and (41) dominate for small and large μ, respectively. Note that $\bar{r}$ is analytic at $\mu/\gamma = \frac{1}{4}$, we thus have no fitness threshold at this point.

Proof of Theorem 4: For a wildtype threshold to occur, $r(x) - g(\mu, x)$ must be maximized at $x = x_{\min}$ for some $\mu > 0$. Therefore, the existence of a wildtype threshold implies an upper bound of $1/\mu_c^-$ on the LHS of (74), which proves the 'only if' part. For the 'if' part, assume that there are sequences x_i in $]x_{\min}, x_{\max}]$ and $\mu_i > 0$ with $\mu_i \to 0$ and $r(x_i) - \mu_i\, g(1, x_i) \geq r(x_{\min}) - \mu_i\, g(1, x_{\min})$ for all i. Let $x_j \to x_\infty$ be a convergent subsequence. Since r and g are assumed to be continuous, we have $r(x_\infty) \geq r(x_{\min})$ and hence $x_\infty = x_{\min}$, since $r(x_{\min})$ is the unique maximum of r by assumption (A1). Thus, we find

$$\frac{g(1, x_j) - g(1, x_{\min})}{r(x_j) - r(x_{\min})} \geq \frac{1}{\mu_j} \to \infty\,, \tag{75}$$

contradicting (74) and thus completing the proof. □

Note that a wildtype threshold will always lead to non-analytic behavior of $\hat{x}(\mu)$ and $\bar{r}(\mu)$ at μ_c^- and is therefore closely related to a fitness threshold. In general, however, it need not show up as a prominent feature with a jump in means or variances. If we have a fitness threshold with a jump in $\hat{x}$ at $\hat{x}(\mu) = x_{\min}$, however, this will also be a wildtype threshold. In a system with a series of thresholds, the wildtype threshold (if it exists) is always the one with the smallest μ_c. An example in shown in Figure 8.

4.3 Degradation thresholds

A far reaching effect of the error threshold is that selection altogether ceases to operate.

Definition 3. *A* degradation threshold *is the smallest mutation rate μ_c^+ above which the population mean fitness is insensitive to any further increase of the mutation rate,*

$$\hat{x}(\mu) = x_{\max}\,, \quad \mu > \mu_c^+\,. \tag{76}$$

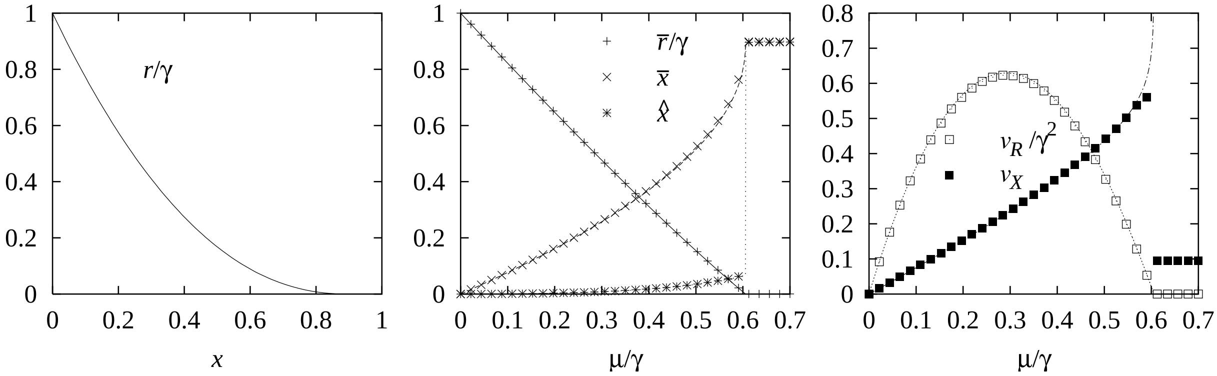

Figure 9: Means (middle) and variances (right) for the biallelic model with asymmetric mutation ($\kappa = 0.8$), $N = 100$ (symbols), and the fitness function $r(x) = \gamma\,(x_{\max} - x)^q/(x_{\max})^q$ with $x_{\max} = (1+\kappa)/2 = 0.9$ and $q = 2.2$ (left). One finds a **degradation threshold** ($\mu_c^+/\gamma \simeq 0.606$), which is also a fitness threshold. As $\hat{r}$ behaves just like $r(\hat{x})$ with a similar accuracy of the approximation, it is not shown here.

Also, the other means and variances then coincide with their values in mutation equilibrium, and the population is degenerate.

Theorem 5. *A degradation threshold occurs if and only if*

$$\lim_{x \nearrow x_{\max}} \frac{r(x) - r(x_{\max})}{g(1,x)} < \infty\,. \tag{77}$$

Proof: For a degradation threshold, $r(x) - g(\mu, x)$ is maximal at $x = x_{\max}$ (where $g(\mu, x_{\max}) = 0$) for any finite $\mu > \mu_c^+$. On the one hand, existence of the threshold implies the criterion with a bound μ_c^+. On the other hand, if we have sequences x_i in $[x_{\min}, x_{\max}[$ and $\mu_i \to \infty$ with $r(x_i) - \mu_i\, g(1, x_i) \geq r(x_{\max})$, we can again choose a convergent subsequence $x_j \to x_\infty$. Since $g(1,x)$ is continuous and $x_{\max}$ is the only zero of g in $[x_{\min}, x_{\max}]$ by (A2), we have $x_\infty = x_{\max}$. As in the wildtype case above, this contradicts (77) and proves the criterion. □

The degradation threshold is related to the fitness threshold in an analogous way as the wildtype threshold above. In particular, we always find non-analytic behavior of $\hat{x}(\mu)$ and $\bar{r}(\mu)$ at μ_c^+, but not necessarily a jump or a kink. However, a fitness threshold with a jump of $\hat{x}(\mu)$ onto $x_{\max}$ is necessarily a degradation threshold. Examples for a degradation threshold are given in Figures 9 and 10.

4.4 Trait thresholds

As stated above, there is usually a kink in the population mean genotype $\bar{x}(\mu)$ at a fitness threshold. The most pronounced change in the equilibrium distribution of x, however, is captured by

Definition 4. *A mutation rate μ_c^x is a* trait threshold *if $\bar{x}$ is discontinuous for $\mu = \mu_c^x$.*

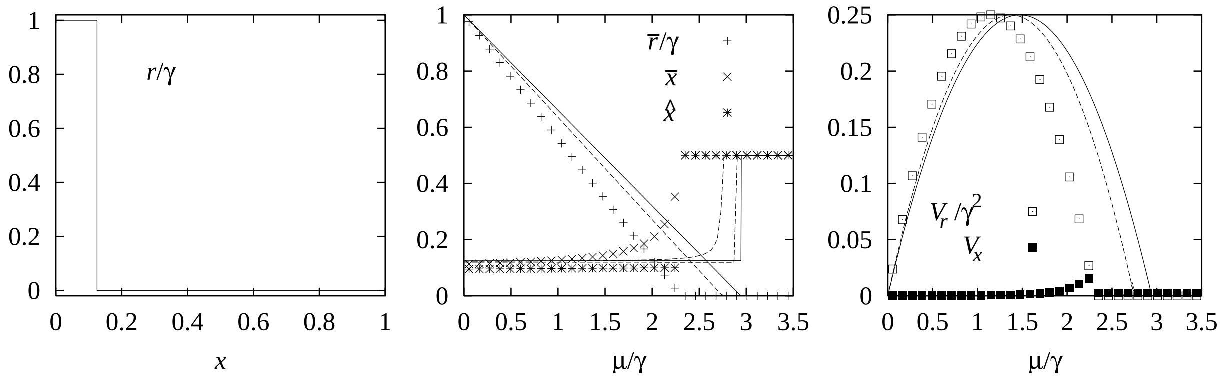

Figure 10: Means (middle) and variances (right) for a model with symmetric mutation ($\kappa = 0$) and **truncation selection**, i.e., $r(x) = \gamma$ for $x \leq \frac{1}{8}$ and $r(x) = 0$ otherwise (left). As in the sharply peaked landscape, cf. Figure 6, one finds a combined fitness, wildtype, degradation, and trait threshold ($\mu_c/\gamma \simeq 2.94$). The variance in fitness follows the different kind of scaling described in Theorem 2. Symbols correspond to $N = 100$, dashed lines to $N = 1000$, and solid lines to the expressions in Section 3.1. As $\hat{r}$ behaves just like $r(\hat{x})$ with similar accuracy, it is not shown here. Note that the deviations of the approximate expressions are somewhat stronger (of order $N^{-1/2}$) for fitness functions with jumps, cf. Section 3.7.

Since a discontinuous change in $\bar{x}$ is usually accompanied by a jump in the *local* mutation rates $u^{\pm}(\bar{x})$ as well as $r'(\bar{x})$, it typically also leads to jumps in v_X and v_R. The mean fitness, however, is not at all affected at such a point (if it does not coincide with a fitness threshold as defined above).

Theorem 6. *A trait threshold occurs if and only if r is not strictly decreasing from $x_{\min}$ to $x_{\max}$.*

Proof: Assume r to be strictly decreasing from $x_{\min}$ to $x_{\max}$. Then, since $\bar{r}(\mu)$ varies continuously from $r(x_{\min})$ to $r(x_{\max})$ as μ changes from 0 to ∞, Theorem 2 implies that $\bar{x}$ is continuous as well, which proves the 'only if' part. Otherwise, if there is some μ such that $\bar{r}(\mu) = r(x_1) = r(x_2)$ with $x_{\min} \leq x_1 < x_2 \leq x_{\max}$, it is obvious from Theorem 2 and the monotonicity of $\bar{r}$ that $\bar{x}$ may not vary continuously from $x_{\min}$ to $x_{\max}$. □

Obviously, any fitness landscape with a trait threshold also fulfills $C > 0$ in Theorem 3 and thus has a fitness threshold, but not vice versa. We have $\mu_c \geq \mu_c^x$, i.e., the jump in $\bar{x}$ in general precedes the fitness transition with the jump in $\hat{x}$, see the example in Figure 11. Trait and fitness thresholds should, therefore, be clearly distinguished. In contrast to the fitness threshold (or a phase transition in physics), the trait threshold is not driven by collective (self-enhancing) action, but only mirrors a simple feature of the fitness function.

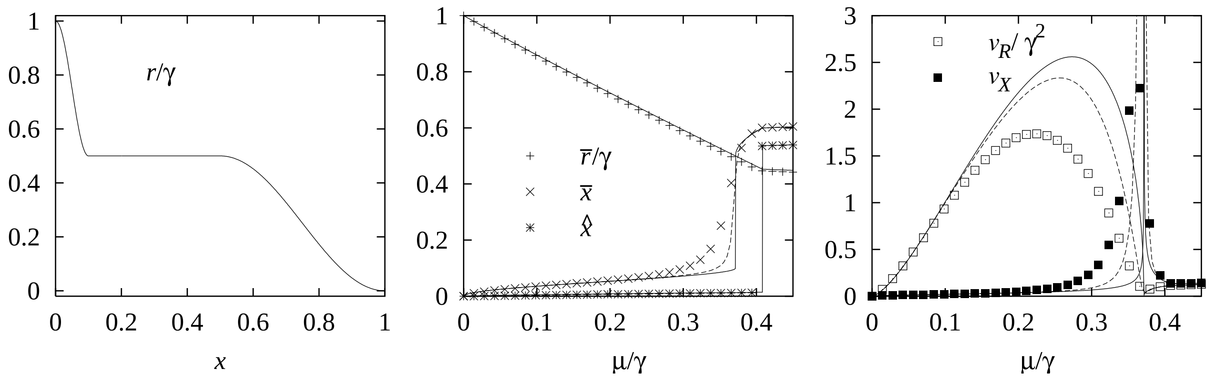

Figure 11: Means (middle) and variances (right) for a model with a fitness function r/γ (left) with an ambiguity for $r(x)/\gamma = 0.5$ and asymmetric mutation ($\kappa = 0.5$). Thus, one finds a **trait threshold** ($\mu_c^x/\gamma \simeq 0.372$), which precedes a fitness threshold ($\mu_c/\gamma \simeq 0.408$), cf. Section 4.4. Symbols correspond to $N = 100$, dashed lines to $N = 500$, and solid lines to the expressions in Section 3.1. As $\hat{r}$ behaves just like $r(\hat{x})$ with similar accuracy, it is not shown here.

II

The continuum-of-alleles model

Mutation–selection models have been described in Section I.1 in a general framework. The rest of Chapter I was then devoted to models with discrete genotypes. This chapter now turns to models in which genotypes are labeled by the elements of some interval $I \subset \mathbb{R}$, which have been introduced by Kimura [Kim65] and are usually referred to as continuum-of-alleles (COA) models. Here, the elements of I refer to genotype classes rather than single genotypes (as in the discrete single-step model (I.3)), defined through their effect on a quantitative trait. The large number of alleles and possible effects justifies the approximation by a continuum. The most important difference to the models of Chapter I is that a different type of mutant distributions is appropriate here.

We will assume the genotype interval to be either compact, most commonly $[0,1]$ or $[-1,1]$, or $\mathbb{R}$ itself. As in Section I.1, the reproduction rates are described by a function $r\colon I \to \mathbb{R}$ and the mutation rates by $u\colon I \times I \to \mathbb{R}_{\geq 0}$, where $u(x,y)$ corresponds to a mutation from type y to type x. We define $u_1\colon I \to \mathbb{R}_{\geq 0}$ as the function of total mutation rates,

$$u_1(x) = \int_I u(y,x)\,\mathrm{d}y \qquad \text{for all } x \in I. \tag{1}$$

Then $u(x,y)/u_1(y)$ is the density function of the distribution of mutant types x, conditioned on a mutation to occur in type y, which happens with rate $u_1(y)$. In correspondence to the time evolution equation (I.1), the equilibrium genotype density p is a solution of

$$\big(r(x) - u_1(x)\big)p(x) + \int_I u(x,y)\,p(y)\,\mathrm{d}y = \lambda\, p(x) \qquad \text{for all } x \in I \tag{2}$$

that fulfills $p \geq 0$ and $\int_I p(x)\,\mathrm{d}x = 1$. For notational convenience later on, we use the symbol λ for the equilibrium mean fitness $\bar{r} = \int_I r(x)\,p(x)\,\mathrm{d}x$.

The first aim of this chapter is to analyze the relationship between the COA model and the discrete model of Chapter I, namely how the former can be approximated by the latter, which is done in Section 2. This is important mainly because it gives a rigorous justification of numerical methods, which are needed since the COA model can, in general, not be treated analytically. Further, it forms a basis on which some results may be transferred from the discrete model.

The second aim is to take first steps towards a simple maximum principle for the equilibrium mean fitness of the COA model, analogous to the central result of Chapter I. In Section 3, some model classes are presented for which explicit upper and lower bounds can be given. In a special case they can be shown to converge towards each other in an appropriate limit and thus indeed establish a simple maximum principle. From numerical examples one may speculate that this is even true for a broader class of models.

The basis for all this are the sufficient criteria for the existence and uniqueness of a solution of the equilibrium condition (2) given by Bürger [Bür88] (with some generalization in [Bür00, Sec. IV.3]). These will be summarized and discussed in some detail in Section 1.

1 General properties

1.1 Operator notation

Let us first put the equilibrium condition (2) in operator notation. Since we are interested in probability densities, we will consider the space of Lebesgue integrable functions on I, $\mathrm{L}^1(I)$, or a subspace thereof, as the underlying function space. We define, for notational brevity,

$$w = u_1 - r \tag{3}$$

and then, for elements f of the function space and all $x \in I$,

$$(Tf)(x) = w(x)f(x)\,, \tag{4}$$

$$(Uf)(x) = \int_I u(x,y)f(y)\,\mathrm{d}y\,, \tag{5}$$

$$A = T - U\,. \tag{6}$$

With this, (2) is equivalent to the eigenvalue equation

$$(A+\lambda)p = 0\,. \tag{7}$$

Being a (non-zero) multiplication operator, T cannot be compact (compare [Mor95, Thm. 2.1]). Strong results like analogs to the Perron–Frobenius theorem, however, are only available for compact, or at least power compact[1], operators, see Schaefer [Sch74, Ch. V]. Therefore one considers the following family of kernel operators:

$$(K_\alpha f)(x) = \int_I k_\alpha(x,y)f(y)\,\mathrm{d}y\,, \tag{8}$$

where

$$k_\alpha(x,y) = \frac{u(x,y)}{w(y)+\alpha}\,. \tag{9}$$

These are, under conditions that will be given shortly, power compact or even compact. Their connection to the operator A from (6) is stated in the following

Lemma 1 (Bürger). *Let T and U be operators in a Banach space X, with U being bounded, T densely defined, i.e., $\overline{D(T)} = X$, and $T+\alpha$ invertible. Then f is an eigenvector of $A = T - U$ with eigenvalue $-\alpha$, i.e., $0 \neq f \in D(A) = D(T)$ and*

$$(A+\alpha)f = 0\,, \tag{10}$$

[1] An operator is said to be *power compact* if one of its powers is compact.

if and only if $g = (T+\alpha)f$ *is an eigenvector of* $K_\alpha = U(T+\alpha)^{-1}$ *with eigenvalue* 1, *i.e.,*

$$(K_\alpha - 1)g = 0\,. \tag{11}$$

PROOF: This is the statement of [Bür88, Prop. 2.1(i)]. From the simple calculation

$$(K_\alpha - 1)g = (U(T+\alpha)^{-1} - 1)(T+\alpha)f = (U - T - \alpha)f = -(A+\alpha)f \tag{12}$$

the formal equivalence of (10) and (11) follows. Since $(T+\alpha)^{-1}$ is bounded, we have $(T+\alpha)^{-1}g \in D(T)$ for all g. □

So, explicitly in our case, the eigenvalue equation (7) is equivalent to

$$(K_\lambda - 1)q = 0 \tag{13}$$

with $q = (T+\lambda)p$.

1.2 Existence and uniqueness of solutions

We follow Bürger [Bür88] to find sufficient conditions for the existence and uniqueness of a solution of (7). To this end, the corresponding kernel operators K_α from (8) are considered.

An important class of bounded kernel operators from $\mathrm{L}^q(I)$ to $\mathrm{L}^p(I)$ $(1 \le p,\, q \le \infty)$ are the Hille–Tamarkin operators, see [Jör70, Sec. 11.3]. Their kernels need to satisfy

$$|K|_{pq} := \|k_1\|_p < \infty \qquad \text{with} \qquad k_1(x) = \|k(x,.)\|_{q'}\,, \tag{14}$$

where $(Kf)(x) = \int_I k(x,y)\, f(y)\,\mathrm{d}y$, $k(x,.)$ denotes the function $y \mapsto k(x,y)$, and q' is the conjugate exponent to q satisfying $\frac{1}{q} + \frac{1}{q'} = 1$, $1 \le q' \le \infty$. The Hille–Tamarkin norm $|.|_{pq}$ turns the set $\mathcal{H}_{pq}(I)$ of all Hille–Tamarkin operators into a Banach space [Jör70, Thm. 11.5]. Here, we are interested in $p = q = 1$, in which case (14) yields

$$|K|_{11} = \int_I \operatorname*{ess\,sup}_{y\in I} |k(x,y)|\,\mathrm{d}x < \infty \tag{15}$$

and K^2 is compact for every $K \in \mathcal{H}_{11}(I)$ [Jör70, Thm. 11.9].

Let us now turn to kernel operators that are power compact, positive, and irreducible. An operator is called *positive* if it maps the set of non-negative functions into itself, for which, in the case of kernel operators, non-negativity of the kernel is necessary and sufficient [Jör70, p. 122]. A kernel operator is *irreducible* if its kernel satisfies [Sch74, Exm. 4 in Sec. V.6]

$$\int_{I\setminus J}\int_J k(x,y)\,\mathrm{d}x\,\mathrm{d}y > 0 \qquad \text{for all measurable } J \subset I \text{ with } |J|,\ |I\setminus J| > 0. \tag{16}$$

Here, $|J|$ denotes the Lebesgue measure of a measurable set J. Then the theorem of Jentzsch [Sch74, Thm. V.6.6], which parallels the Perron–Frobenius theorem for matrices, states that the spectral radius is an algebraically simple eigenvalue with an (up to normalization) unique positive eigenfunction (i.e., strictly positive a.e.[2]) and the only eigenvalue with a positive eigenfunction.

In our case, the following requirements are sufficient for the K_α to be Hille–Tamarkin operators [Bür88, Sec. 3].

(U1) u is non-negative and measurable.
(U2) $u_1(x)$ from (1) exists for a.e. $x \in \mathbb{R}$ and u_1 is essentially bounded, $u_1 \in \mathrm{L}^\infty(I)$.
(T1) $w = u_1 - r$ is measurable and satisfies $\operatorname{ess\,inf}_{x\in I} w(x) = 0$. (The latter can be achieved, without loss of generality, by adding a suitable constant to r.)
(T2) $(w+1)^{-1} \in \mathrm{L}^\infty(I)$ is then already a consequence of (T1).
(U4) $\int_I \operatorname{ess\,sup}_{y\in I} u(x,y)/(w(y)+\alpha)\,\mathrm{d}x < \infty$ for one (and then for all) $\alpha > 0$.

For $\alpha > 0$, K_α is irreducible if U is [Bür88, proof of Thm. 2.2(c)], i.e.,

$$\int_{I\setminus J}\int_J u(x,y)\,\mathrm{d}x\,\mathrm{d}y > 0 \qquad \text{for all measurable } J \subset I \text{ with } |J|,\ |I\setminus J| > 0. \tag{17}$$

To keep the equilibrium genotype distribution from having atoms, we assume, similar to [Bür00, cond. 3” in Sec. IV.3], that there is a set $J \subset I$ with positive measure, in which w takes its infimum, $\operatorname{ess\,inf}_{x\in J} w(x) = 0$, such that

$$\operatorname*{ess\,inf}_{x,y\in J} u(x,y) \int_J (w(x))^{-1}\,\mathrm{d}x > 1 \tag{18}$$

or the integral diverges.

Putting everything together, we have the following

Theorem 1 (Bürger). *Under the above conditions,* (2) *has a unique positive solution* $p \in \mathrm{L}^1(I)$ *with* $\|p\|_1 = 1$*, for which* $\lambda > 0$ *is the largest spectral value of* $-A$ *from* (6).

Proof: See the above, [Bür88, Thm. 3.5], and [Bür00, Sec. IV.3]. □

Another result that will be needed in the sequel is

Lemma 2 [Bür88, Lems. 1–3, Thm. 2.2(ii)]. *Under the above conditions, the spectral radius* $\rho(K_\alpha)$ *is, as a function of* α*, strictly decreasing and satisfies* $\rho(K_\lambda) = 1$ *and* $\lim_{\alpha\to\infty}\rho(K_\alpha) = 0$*. Thus,* $\rho(K_\alpha) < 1$ *implies* $\alpha > \lambda$ *and* $\rho(K_\alpha) > 1$ *implies* $\alpha < \lambda$.

2 Discretization

In this section, we will consider two methods to approximate compact kernel operators by operators of finite rank, i.e., operators with a finite-dimensional image. Under certain additional restrictions, these techniques carry over to non-compact operators like A from (6). One, the Nyström method, is applicable to continuous functions r and u on compact intervals I and involves sampling[3] of these functions. The other, the Galerkin method,

[2] The abbreviation ‘a.e.’ stands for ‘almost every’ or ‘almost everywhere’ and means that the set at which the condition it refers to is not fulfilled has zero (Lebesgue) measure.

[3] The term *sampling* is used in the meaning also used in signal processing: Instead of a continuous function one considers its values at a (properly chosen) finite set of points.

is based on projections to finite-dimensional subspaces and works—in principle—for a broad class of measurable kernels. In order to get an approximation for A, however, one has to make quite strong assumptions, e.g., that the functions r and u are, in some sense, uniformly continuous. Then, it turns out, the local averaging in the projection process can be replaced by sampling again (if an additional condition is fulfilled).

As the case of a compact genotype interval is simpler, it will be treated first. Throughout this section we will assume that the criteria from Section 1.2 are satisfied, namely (U1), (U2), (U4), (T1), (T2), (17), and (18).

2.1 Compact genotype interval

Let the genotype interval I be compact and $\mathrm{C}(I)$ denote the Banach space of bounded, continuous functions equipped with the supremum norm $\|f\|_\infty = \sup_{x\in I} |f(x)|$. We consider operators K of the form

$$(Kf)(x) = \int_I k(x,y) f(y) \, \mathrm{d}y \qquad \text{for all } x \in I \tag{19}$$

with a continuous kernel $k \colon I \times I \to \mathbb{R}$. First note these two basic results:

Proposition 1. *An operator K of the form* (19) *maps $\mathrm{L}^1(I)$ into $\mathrm{C}(I) \subset \mathrm{L}^1(I)$.*

Proof: We follow the proof of [Eng97, Thm. 2.1], where this is shown for $\mathrm{L}^2(I)$, which, since I is compact, is a subspace of $\mathrm{L}^1(I)$. Let $f \in \mathrm{L}^1(I)$ and x, $\xi \in I$ be given. Then

$$|(Kf)(x) - (Kf)(\xi)| \le \int_I |k(x,y) - k(\xi,y)| \, |f(y)| \, \mathrm{d}y \le \sup_{y\in I} |k(x,y) - k(\xi,y)| \, \|f\|_1 \, . \tag{20}$$

Due to the uniform continuity of k in $I \times I$, we have

$$\lim_{\xi\to x} \sup_{y\in I} |k(x,y) - k(\xi,y)| = 0 \, , \tag{21}$$

from which the continuity of Kf follows. □

Proposition 2. *An operator K of the form* (19) *is compact from $\mathrm{C}(I)$ or $\mathrm{L}^1(I)$ to either of the two spaces.*

Proof: Follow the proof of [Eng97, Thm. 2.10] (or [Lan93, XVII.4]), where this is shown for $\mathrm{L}^2(I) \subset \mathrm{L}^1(I)$, and use Hölder's inequality whenever the Cauchy–Schwarz inequality is used. Alternatively, see [Sch74, Exm. 3 in Sec. IV.10]. □

Thus, if in our case the functions r and u are continuous, also the kernel k_α is, for every $\alpha > 0$. It then follows from Proposition 1 that the equilibrium genotype density p is continuous as well. Therefore we can restrict our attention to $\mathrm{C}(I)$ in our quest for a solution of the eigenvalue equation (7). This makes the Nyström method applicable as a discretization procedure, which will be presented now.

2.1.1 The Nyström method

The Nyström method is based on quadratures, which are used for numerical integration, compare Kress [Kre99, Ch. 12]. We will use this (slightly restricted)

Definition 1. *A* quadrature rule Q_n *is a mapping of the form*

$$Q_n \colon \mathrm{C}(I) \to \mathbb{R}, \qquad f \mapsto Q_n f = \sum_{k=1}^{N_n} \alpha_{n,k} f(t_{n,k}), \tag{22}$$

with $n \in \mathbb{N}$, $N_n \in \mathbb{N}$, quadrature points $t_{n,k} \in I$, *and* quadrature weights $\alpha_{n,k} > 0$, *for* $k \in \mathcal{N}_n := \{1, \ldots, N_n\}$. *A sequence of quadrature rules, or simply a* quadrature, (Q_n) *is said to be* convergent *if*

$$Q_n f \to Q f \qquad \textit{for all } f \in \mathrm{C}(I), \tag{23}$$

where Q *is the linear functional that maps each* $f \in \mathrm{C}(I)$ *to its integral,* $Qf = \int_I f(x)\,\mathrm{d}x$.

Another notion that is important for the Nyström method is the collectively compact convergence of operators. The standard reference for this matter is Anselone's book [Ans71].

Definition 2. *A sequence* (K_n) *of (compact) operators in a Banach space* X *is* collectively compact *provided that the set* $\{K_n B : n \in \mathbb{N}\}$ *is relatively compact (i.e., its closure is compact) for every bounded set* $B \subset X$. *If furthermore the sequence converges pointwise to an operator* K *one speaks of* collectively compact convergence, *in symbols* $K_n \xrightarrow{\mathrm{cc}} K$.

As a direct consequence of this definition, K is compact (as well as all K_n).

The central result for the Nyström method is

Theorem 2. *Let* K *be a compact kernel operator of the form* (19) *whose eigenvalue equation*

$$(K - \nu) g = 0 \tag{24}$$

is to be approximated. To this end, let $(Q_n)_{n \in \mathbb{N}}$ *be a convergent quadrature with the notation as in Definition* 1. *A* complete discretization *is given by the* $N_n \times N_n$ *matrices* $\boldsymbol{K}_n$ *with entries*

$$K_{n,k\ell} = \alpha_{n,\ell}\, k(t_{n,k}, t_{n,\ell}), \tag{25}$$

a partial discretization *by means of the operators* K_n *on* $\mathrm{C}(I)$ *with*

$$(K_n f)(x) = \sum_{k=1}^{N_n} \alpha_{n,k}\, k(x, t_{n,k}) f(t_{n,k}) = Q_n(k(x, .)f). \tag{26}$$

Consider the corresponding eigenvalue equations

$$(\boldsymbol{K}_n - \nu_n)\boldsymbol{g}_n = 0 \qquad \textit{and} \qquad (K_n - \nu_n) g_n = 0, \tag{27}$$

where $\boldsymbol{g}_n$ *is an* N_n*-dimensional vector with components* $g_{n,k}$, *and* $g_n \in \mathrm{C}(I)$. *Then, under the above conditions the following statements are true:*

(a) *Both eigenvalue equations in* (27) *are equivalent and connected via*

$$g_n(x) = \sum_{k=1}^{N_n} \alpha_{n,k}\, k(x, t_{n,k}) g_{n,k}\,. \tag{28}$$

(b) *For every* $\nu \neq 0$ *from* (24) *there is a sequence* (ν_n) *of eigenvalues of* (27) *such that* $\nu_n \to \nu$ *as* $n \to \infty$*. Conversely, every non-zero limit point of any sequence* (ν_n) *of eigenvalues of* (27) *is an eigenvalue of* (24)*.*

(c) *Every bounded sequence* (g_n) *of eigenfunctions of* (27) *associated with eigenvalues* $\nu_n \to \nu \neq 0$ *contains a convergent subsequence; the limit of any convergent subsequence* $(g_{n_i})_i$ *is an eigenfunction of* (24) *associated with the eigenvalue* ν *(unless the limit is zero).*

Proof: (a) is the statement of [Kre99, Thm. 12.7] or [Eng97, Lem. 3.15]. (b) and (c) rely on the collectively compact convergence $K_n \xrightarrow{\text{cc}} K$, which is shown, e.g., in [Ans71, Props. 2.1, 2.2], [Kre99, Thm. 12.8], or [Eng97, Thm. 3.22]. The statements then follow from [Ans71, Thms. 4.11, 4.17]. □

With respect to the discretization procedure we are aiming for, we will restrict ourselves to quadratures that allow for disjoint partitions of I with intervals $I_{n,k}$, i.e., $I_{n,k} \cap I_{n,\ell} \neq \emptyset$ and $\bigcup_{k=0}^{N_n} I_{n,k} = I$, such that $t_{n,k} \in I_{n,k}$ and $|I_{n,k}| = \alpha_{n,k}$ (with $k \in \mathcal{N}_n$). For such quadratures it is easy to see that[4]

$$\|Q_n\| = \sum_{k=1}^{N_n} \alpha_{n,k} = |I| \tag{29}$$

and that the partitions are unique (up to the boundary points of the intervals). Furthermore we have

Lemma 3. *Let* (Q_n) *be a convergent quadrature that allows for partitions of* I *as described above. Then* $\lim_{n\to\infty} \max_{k\in\mathcal{N}_n} |I_{n,k}| = 0$.

Proof: Assume the contrary. Then there are an $\varepsilon > 0$ and sequences $(n_i)_i$ and $(k_i)_i$ with $\lim_{i\to\infty} n_i = \infty$ and $|I_{n_i,k_i}| \geq \varepsilon$. Due to the compactness of I, these can be chosen such that $\lim_{i\to\infty} t_{n_i,k_i} =: t$ exists. Consider $f(x) = \max\{1-2|x-t|/\varepsilon, 0\}$, which is a continuous function. For this we have $Qf \leq \varepsilon/2$, but $\lim_{i\to\infty} Q_{n_i} f \geq \varepsilon \lim_{i\to\infty} f(t_{n_i,k_i}) = \varepsilon$, which contradicts the convergence of the quadrature (23). □

2.1.2 Application to the COA model

In our case of the COA model with a compact interval I and continuous functions r and u, the complete discretization is given by the following $N_n \times N_n$ matrices:

$$T_{n,k\ell} = \delta_{k\ell}\, w(t_{n,k}) \geq 0\,, \tag{30}$$

$$U_{n,k\ell} = \alpha_{n,\ell}\, u(t_{n,k}, t_{n,\ell}) \geq 0\,, \tag{31}$$

$$\boldsymbol{A}_n = \boldsymbol{T}_n - \boldsymbol{U}_n, \qquad \boldsymbol{K}_{\alpha,n} = \boldsymbol{U}_n(\boldsymbol{T}_n + \alpha)^{-1} \quad \text{for } \alpha > -\min_{k\in\mathcal{N}_n} w(t_{n,k}). \tag{32}$$

[4] If not noted otherwise, the following convention for operator norms is used. If an operator maps a space X into itself, we denote its norm by the same symbol as the norm of X, e.g., $\|.\|_X$, or $\|.\|_1$ for L^1; in all other cases the unornamented symbol $\|.\|$ is used.

The eigenvalue equations to be solved are

$$(\boldsymbol{A}_n + \lambda_n)\boldsymbol{p}_n = 0 \qquad \text{with } \boldsymbol{p}_n > 0. \tag{33}$$

Here, $-\boldsymbol{A}_n + c$ is positive with a suitable constant c. We further have to assume that the $\boldsymbol{A}_n$ are irreducible (which might not be the case for special choices of the $t_{n,k}$, e.g., if $u_1(t_{n,k}) = 0$ for some k). Then, due to the Perron–Frobenius theorem, there exist (up to normalization) unique positive $\boldsymbol{p}_n$ belonging to the eigenvalues $-\lambda_n = -\rho(-\boldsymbol{A}_n + c) + c$, where $\rho(\boldsymbol{M})$ denotes the spectral radius of a matrix $\boldsymbol{M}$, cf. the end of Section I.2.1. With $\boldsymbol{q}_n = (\boldsymbol{T}_n + \lambda_n)\boldsymbol{p}_n$ also the eigenvalue equations

$$(\boldsymbol{K}_{\lambda_n,n} - 1)\boldsymbol{q}_n = 0 \tag{34}$$

are solved (and vice versa), cf. Lemma 1.

Both $\boldsymbol{K}_{\lambda_n,n}$ and $\boldsymbol{q}_n$ can be embedded into $\mathrm{C}(I)$ as described by (26) and (28). Then, with Theorem 2, one might conclude the convergence $\|q_n - q\|_\infty \to 0$. In the end, however, we are interested in the population vectors $\boldsymbol{p}_n$ and their convergence to the density p. It might be easiest to interpret the vectors $\boldsymbol{p}_n$ as point measures on I. But then the best one can hope for is weak convergence since the set of point measures is closed under the total variation norm. It will turn out that we can indeed achieve norm convergence if we embed the $\boldsymbol{p}_n$ into $\mathrm{L}^1(I)$ the following way. We choose a disjoint partition of I as above and let

$$p_n = \sum_{k=1}^{N_n} p_{n,k} 1_{I_{n,k}}\,, \tag{35}$$

where 1_J denotes the characteristic function of a set J. (Note that $p_{n,k}$ denotes the k-th component of $\boldsymbol{p}_n \in \mathbb{R}^{N_n}$, whereas p_n is an L^1 function.) Thus the $\boldsymbol{p}_n$ can be interpreted as probability densities on I, if we normalize them such that $\|p_n\|_1 = 1$. This is most easily expressed using the induced norm $\|\boldsymbol{f}\|_{(n)} := \sum_{k=1}^{N_n} \alpha_{n,k}|f_k|$ on $\mathbb{R}^{N_n}$. Convergence in total variation then corresponds to $\|p_n - p\|_1 \to 0$ [Rud86, Thm. 6.13]. One can also define operator analogs of the $\boldsymbol{A}_n$ by

$$A_n f = \sum_{k=1}^{N_n} 1_{I_{n,k}} \sum_{\ell=1}^{N_n} A_{n,k\ell} \frac{1}{|I_{n,\ell}|} \int_{I_{n,\ell}} f(x)\,\mathrm{d}x\,, \tag{36}$$

for which $(A_n + \lambda_n)p_n = 0$ holds. These, however, will not be used in the sequel.[5]

2.1.3 Convergence of eigenvalues and eigenvectors

We now come to prove the main approximation result:

Theorem 3. *With the notation and assumptions from Sections* 1.1 *and* 2.1.2*,*

(a) $\lim_{n\to\infty} \lambda_n = \lambda > 0$ *and*

[5] The A_n are of finite rank and thus compact. Therefore, $A_n \to A$ can hold neither in the norm nor in the collectively compact sense, since then A would be compact as well.

(b) $\lim_{n\to\infty} \|p_n - p\|_1 = 0$, *i.e., the probability measures corresponding to these densities converge in total variation.*

The idea of the proof of part (a) is as follows. In the following two lemmas, we first determine an upper and a lower bound for the λ_n and conclude that there is a convergent subsequence. Then we show that every convergent subsequence converges to λ and hence the sequence itself.

Lemma 4. *There is a constant $M > 0$ such that $|\lambda_n| \leq M$ for all $n \in \mathbb{N}$.*

Proof: Using (30) and (31), one checks

$$|\lambda_n| = \frac{\|\lambda_n \boldsymbol{p}_n\|_{(n)}}{\|\boldsymbol{p}_n\|_{(n)}} = \frac{\|\boldsymbol{A}_n \boldsymbol{p}_n\|_{(n)}}{\|\boldsymbol{p}_n\|_{(n)}} \leq \sup_{\|\boldsymbol{f}\|_{(n)}=1} \sum_{k=1}^{N_n} \alpha_{n,k} \left| \sum_{\ell=1}^{N_n} (T_{n,k\ell} - U_{n,kl}) f_\ell \right|$$
$$\leq \max_k w(t_{n,k}) + \max_{k,\ell} u(t_{n,k}, t_{n,\ell}) \sum_{k=1}^{N_n} \alpha_{n,k} \leq \|w\|_\infty + \|u\|_{C(I\times I)} \sup_m \|Q_m\| =: M > 0\,.$$

Here, $\|Q_m\| = |I|$ due to (29). More generally, for any convergent quadrature, according to the theorem of Banach–Steinhaus, $\sup_m \|Q_m\| < \infty$, compare [Rud91, Thm. 2.5]. ☐

Lemma 5. $\liminf_{n\to\infty} \lambda_n > 0$.

Proof: We start by following Bürger [Bür00, p. 134] and show that the spectral radius $\rho(K_\alpha)$ is larger than 1 for sufficiently small $\alpha > 0$, from which then $\lambda > \alpha > 0$ follows by Lemma 2. Let J be the interval from (18). Then we have

$$(K_\alpha 1_J)(x) = \int_J \frac{u(x,y)}{w(y)+\alpha}\,\mathrm{d}y \geq 1_J(x) \operatorname*{ess\,inf}_{x',y'\in J} u(x',y') \int_J (w(y)+\alpha)^{-1}\,\mathrm{d}y \tag{37}$$

and thus

$$\|K_\alpha{}^m\|_1{}^{1/m} \geq \operatorname*{ess\,inf}_{x,y\in J} u(x,y) \int_J (w(y)+\alpha)^{-1}\,\mathrm{d}y \qquad \text{for all } m \in \mathbb{N}, \tag{38}$$

which implies that the spectral radius satisfies

$$\rho(K_\alpha) \geq \operatorname*{ess\,inf}_{x,y\in J} u(x,y) \int_J (w(y)+\alpha)^{-1}\,\mathrm{d}y. \tag{39}$$

The RHS is, as a function of α, strictly decreasing. Thus, as a consequence of B. Levi's monotone convergence theorem [Hew69, Thm. III.12.22], also

$$\lim_{\alpha\searrow 0} \rho(K_\alpha) \geq \operatorname*{ess\,inf}_{x,y\in J} u(x,y) \int_J (w(y))^{-1}\,\mathrm{d}y > 1 \tag{40}$$

according to (18) (including divergence of both sides).

Now we choose $\alpha > 0$ such that the RHS of (39) is larger than or equal to $1+\varepsilon$, with a sufficiently small $\varepsilon > 0$. Furthermore, we pick, according to the convergence of the quadrature, an n_0 with $\operatorname{ess\,inf}_{x,y\in J} u(x,y)|Q_n(w+\alpha)^{-1} - Q(w+\alpha)^{-1}| < \varepsilon/2$ for all $n \geq n_0$. This way

$$\begin{aligned}(K_{\alpha,n}1_J)(x) &= Q_n\big(u(x,.)(w+\alpha)^{-1}1_J\big) \geq 1_J(x) \operatorname*{ess\,inf}_{x',y'\in J} u(x',y')Q_n(w+\alpha)^{-1} \\ &\geq 1_J(x)\left(\operatorname*{ess\,inf}_{x',y'\in J} u(x',y')Q(w+\alpha)^{-1} - \frac{\varepsilon}{2}\right) \geq 1_J(x)\left(1+\frac{\varepsilon}{2}\right),\end{aligned} \tag{41}$$

and hence, by Lemma 2, $\lambda_n > \alpha > 0$ for all $n \geq n_0$, from which the claim follows. □

Proof of Theorem 3(a): According to Lemmas 4 and 5, the sequence $(\lambda_n)_n$ has a convergent subsequence $(\lambda_{n_i})_i$ with limit $\lambda' \in]0, M]$. Consider $(K_{\lambda'}f)(x) = Q(k_{\lambda'}(x,.)f)$ as well as $(K_nf)(x) := (K_{\lambda_n,n}f)(x) = Q_n(T+\lambda_n)^{-1}(T+\lambda')(k_{\lambda'}(x,.)f)$.

We first show that the 'distorted' quadrature $\tilde{Q}_{n_i} = Q_{n_i}(T+\lambda_{n_i})^{-1}(T+\lambda')$ is convergent. To this end, note that, for i_0 large enough, such that $\inf_{j\geq i_0}\lambda_{n_j} > 0$, and $i \geq i_0$,

$$\begin{aligned}\|(T+\lambda_{n_i})^{-1}(T+\lambda) - 1\|_\infty &= \sup_{\|f\|_\infty\leq 1}\left\|\frac{w+\lambda}{w+\lambda_{n_i}}f - f\right\|_\infty \\ &\leq \|(w+\inf_{j\geq i_0}\lambda_{n_j})^{-1}\|_\infty\,|\lambda-\lambda_{n_i}|\,\|f\|_\infty \to 0\,.\end{aligned} \tag{42}$$

Then, since (Q_n) is convergent by assumption, we have, for all $f \in \mathrm{C}(I)$,

$$\|(\tilde{Q}_{n_i} - Q)f\|_\infty \leq \|Q_{n_i}((T+\lambda_{n_i})^{-1}(T+\lambda') - 1)f\|_\infty + \|(Q_{n_i} - Q)f\|_\infty \to 0, \tag{43}$$

where the first term vanishes in the limit due to $\sup_m \|Q_m\| < \infty$ and (42).

With this it follows from Theorem 2 that $\rho(K_n) = 1$ is also an eigenvalue of $K_{\lambda'}$ going with a non-negative eigenfunction. The latter is even a.e. positive since, due to the irreducibility (17) of $K_{\lambda'}$, there cannot be a set with positive measure on which a non-negative eigenfunction vanishes.[6] But since, according to Theorem 1, there is, up to normalization, only one positive eigenfunction, we have $\lambda' = \lambda$. Therefore every convergent subsequence of $(\lambda_n)_n$ converges to λ, and thus, due to the boundedness, also the sequence itself. □

The strategy for the proof of part (b) of Theorem 3 is as follows. We first show that the norms $\|p_n\|_1$ and $\|q_n\|_\infty$ are (ultimately) 'equivalent'. Thus, as $\|p_n\|_1 = 1$, the sequence $(\|q_n\|_\infty)$ has a convergent subsequence. We then demonstrate that convergence of the norms of a subsequence $(q_{n_i})_i$ implies the convergence of the subsequence itself. Finally, the claim is proven by showing that every convergent subsequence of (q_n) converges to q from (13) and hence the sequence itself.

Lemma 6. *There are constants c_1, $c_2 > 0$ and an $n_0 \in \mathbb{N}$ such that*

$$0 < c_1\|p_n\|_1 \leq \|q_n\|_\infty \leq c_2\|p_n\|_1 \qquad \textit{for every } n \geq n_0. \tag{44}$$

[6] Let $\tilde{q}$ be the eigenfunction and $J = \{x : \tilde{q}(x) > 0\}$ with $|J| > 0$. Assume $|J| < |I|$. Then, for $x \in I\backslash J$, we have $0 = \tilde{q}(x) = \int_J k_{\lambda'}(x,y)\tilde{q}(y)\,\mathrm{d}y$, which implies, for a.e. $y \in J$, that $k_{\lambda'}(x,y)\tilde{q}(y) = 0$ and thus $u(x,y) = 0$, contradicting (17).

PROOF: On the one hand, due to (34),

$$\|q_n\|_\infty = \sup_{x\in I}\sum_{\ell=1}^{N_n}\alpha_{n,\ell}\,k_{\lambda_n}(x,t_{n,\ell})\,q_{n,\ell} \ge \max_{k\in\mathcal{N}_n}\sum_{\ell=1}^{N_n}\alpha_{n,\ell}\,k_{\lambda_n}(t_{n,k},t_{n,\ell})\,q_{n,\ell} = \max_{k\in\mathcal{N}_n} q_{n,k}\,. \tag{45}$$

Choose n_0 according to Lemma 5 such that $\lambda_n \ge \alpha$ for some $\alpha > 0$ and every $n \ge n_0$. Let $c_1 := |I|^{-1}\alpha > 0$. Then we have, for every $n \ge n_0$, recalling $\boldsymbol{q}_n = (\boldsymbol{T}_n + \lambda_n)\boldsymbol{p}_n$,

$$\begin{aligned} 0 < c_1\|p_n\|_1 = c_1\sum_{k=1}^{N_n}\alpha_{n,k}\,p_{n,k} &= c_1\sum_{k=1}^{N_n}\alpha_{n,k}(T_{n,kk}+\lambda_n)^{-1}q_{n,k} \\ &\le c_1|I|\,\alpha^{-1}\max_{k\in\mathcal{N}_n} q_{n,k} \le \|q_n\|_\infty\,. \end{aligned} \tag{46}$$

On the other hand, using $\boldsymbol{K}_{\lambda_n,n}\boldsymbol{q}_n = \boldsymbol{U}_n(\boldsymbol{T}_n+\lambda_n)^{-1}(\boldsymbol{T}_n+\lambda_n)\boldsymbol{p}_n = \boldsymbol{U}_n\boldsymbol{p}_n$,

$$\|q_n\|_\infty = \sup_{x\in I}\sum_{\ell=1}^{N_n}\alpha_{n,\ell}\,u(x,t_{n,\ell})\,p_{n,\ell} \le \|u\|_{\mathrm{C}(I\times I)}\|\boldsymbol{p}_n\|_{(n)} =: c_2\|p_n\|_1\,, \tag{47}$$

which completes the proof. □

Lemma 7. *If* $\|q_n\|_\infty \to c$ *holds, then also* $q_n \to c\,q/\|q\|_\infty$ *with* q *from* (13).

PROOF: Due to the collective compactness of (K_n), see [Kre99, Thm. 12.8] or [Eng97, Thm. 3.22], the $q_n = K_n q_n$ are contained in a compact set. Hence there is a convergent subsequence $(q_{n_i})_i$, whose limit we denote by $\tilde q$, with $\|\tilde q\|_\infty = c$ and $\tilde q \ge 0$. Consider

$$\|\tilde q - K_\lambda\tilde q\|_\infty \le \|\tilde q - K_{n_i}q_{n_i}\|_\infty + \|K_{n_i}\|_\infty\|q_{n_i}-\tilde q\|_\infty + \|K_{n_i}\tilde q - K_\lambda\tilde q\|_\infty. \tag{48}$$

As $i\to\infty$, the first term vanishes due to $K_{n_i}q_{n_i} = q_{n_i}$, the second since, according to the theorem of Banach–Steinhaus, $\sup_n\|K_n\|_\infty < \infty$, and the third due to the pointwise convergence of K_{n_i} towards K_λ. Thus, $\tilde q$ is an eigenfunction of K_λ to the eigenvalue 1 and, according to the theorem of Jentzsch [Sch74, Thm. V.6.6], unique up to normalization, and therefore $\tilde q = c\,q/\|q\|_\infty$. Since this is true for all convergent subsequences, the claim follows (again due to the collective compactness). □

PROOF OF THEOREM 3(b): With our choice $\|p_n\|_1 = 1$ there is, due to Lemma 6, a subsequence $(q_{n_i})_i$ such that $\|q_{n_i}\|_\infty$ converges. Let us denote the limit by c. Hence, according to Lemma 7, also $q_{n_i} \to cq/\|q\|_\infty =: \tilde q$ holds. Now consider

$$\begin{aligned} \|p_{n_i} - (T+\lambda)^{-1}\tilde q\|_1 &= \sum_{k=1}^{N_{n_i}}\int_{I_{n_i,k}}\left|p_{n_i,k} - (w(x)+\lambda)^{-1}\tilde q(x)\right|\mathrm{d}x \\ &\le \sum_{k=1}^{N_{n_i}}|I_{n_i,k}|\sup_{x\in I_{n_i,k}}\left|p_{n_i,k} - (w(x)+\lambda)^{-1}\tilde q(x)\right| \\ &\le |I|\max_k\sup_{x\in I_{n_i,k}}\left|(w(t_{n_i,k})+\lambda_{n_i})^{-1}q_{n_i,k} - (w(x)+\lambda)^{-1}\tilde q(x)\right| \\ &\le |I|\max_k\left|(w(t_{n_i,k})+\lambda_{n_i})^{-1} - (w(t_{n_i,k})+\lambda)^{-1}\right|q_{n_i}(t_{n_i,k}) + \\ &\quad |I|\max_k(w(t_{n_i,k})+\lambda)^{-1}\left|q_{n_i}(t_{n_i,k}) - \tilde q(t_{n_i,k})\right| + \\ &\quad |I|\max_k\sup_{x\in I_{n_i,k}}\left|(w(t_{n_i,k})+\lambda)^{-1}\tilde q(t_{n_i,k}) - (w(x)+\lambda)^{-1}\tilde q(x)\right|\,. \end{aligned} \tag{49}$$

The first term is bounded from above by

$$|I|\,|\lambda - \lambda_{n_i}|\,\|(w + \inf_{m \geq n_0} \lambda_m)^{-1}(w+\lambda)^{-1}\|_\infty \sup_i \|q_{n_i}\|_\infty\,, \tag{50}$$

for $n_i \geq n_0$ with sufficiently large n_0, and vanishes for $i \to \infty$ because of $\lambda_n \to \lambda$ and the boundedness of $\|q_{n_i}\|_\infty$. The second term vanishes due to the uniform convergence of the q_{n_i} towards $\tilde{q}$, and the third due to the uniform continuity of $(w+\lambda)^{-1}\tilde{q}$ and Lemma 3. With this we have $p_{n_i} \to (T+\lambda)^{-1}\tilde{q}$ in $\mathrm{L}^1(I)$, and hence $\|(T+\lambda)^{-1}\tilde{q}\|_1 = 1$, from which $c = \|q\|_\infty$ follows. Therefore, each convergent subsequence of $(q_n)_n$ converges to q and thus, as in the proof of Lemma 7, the sequence itself. Together with what has just been shown the claim follows. □

2.2 Unbounded genotype interval

Now we assume the genotypes to be taken from $I = \mathbb{R}$ and the functions r and u to be continuous. It will be one aim of this section to analyze what further conditions have to be imposed in order to allow for a discretization procedure similar to the one in the previous section. In order to do so, we start by a summary of the relevant theory.

2.2.1 The Galerkin method

In the Galerkin method, an approximation of compact operators is achieved using projections to finite-dimensional subspaces. This method has been reviewed, e.g., by Krasnosel'skii et al. [Kra72, Sec. 18]. The results needed in the sequel are collected in the following

Theorem 4. *Let K be a compact linear operator on the Banach space Y. Consider the eigenvalue equation*

$$(K - \nu)g = 0\,, \tag{51}$$

which is to be approximated. To this end, let (Y_n) be a sequence of closed subspaces of Y with bounded projections P_n onto them. On these subspaces, let the compact linear operators K_n be defined, together with the eigenvalue equations

$$(K_n - \nu_n)g_n = 0\,. \tag{52}$$

Assume that

$$\|K_n - P_n K\|_{Y_n} \to 0\,, \quad \|K - P_n K\|_Y \to 0 \qquad \text{as } n \to \infty. \tag{53}$$

Then the following statements are true:

(a) *For every $\nu \neq 0$ from* (51) *there is a sequence (ν_n) of eigenvalues of* (52) *such that $\nu_n \to \nu$ as $n \to \infty$. Conversely, every non-zero limit point of any sequence (ν_n) of eigenvalues of* (52) *is an eigenvalue of* (51).

(b) *Every bounded sequence (g_n) of eigenvectors of* (52) *associated with eigenvalues $\nu_n \to \nu \neq 0$ contains a convergent subsequence; the limit of any convergent subsequence $(g_{n_i})_i$ is an eigenvector of* (51) *associated with the eigenvalue ν* (*unless the limit is zero*).

PROOF: See [Kra72, Thms. 18.1, 18.2], where also certain cases of unbounded projections are treated. □

A sufficient condition for the validity of the second assumption in (53) is given by the following

Proposition 3. *Let X be a normed space, Y a Banach space, and K a compact linear operator from X into Y, which is to be approximated. For bounded linear operators $P_n\colon Y \to Y$ $(n \in \mathbb{N})$ with $P_n \to \mathbb{1}$ pointwise for $n \to \infty$ we have $\|P_n K - K\| \to 0$.*

PROOF: We follow Werner [Wer00, Thm. II.3.5] (or Engl [Eng97, Rem. 2.8]). Let U denote the unit ball in X and $A = \overline{K(U)} \subset Y$ the completion of its image under K, which is compact due to the compactness of K. Consider the expression

$$\limsup_{n\to\infty} \|P_n K - K\| = \limsup_{n\to\infty} \sup_{x\in U} \|P_n K x - Kx\| \leq \limsup_{n\to\infty} \sup_{y\in A} \|(P_n - \mathbb{1})y\| . \tag{54}$$

Then, for arbitrary $\varepsilon > 0$, there exists a finite ε-net $\{y_1, \ldots, y_k\} \subset A$, i.e., for all $y \in A$ we have $\inf_{1\leq i\leq k} \|y - y_i\| < \varepsilon$. Therefore,

$$\begin{aligned} \sup_{y\in A} \|(P_n - \mathbb{1})y\| &\leq \sup_{1\leq i\leq k} \|(P_n - \mathbb{1})y_i\| + \|P_n - \mathbb{1}\| \inf_{1\leq i\leq k} \|y - y_i\| \\ &\leq \sup_{1\leq i\leq k} \|(P_n - \mathbb{1})y_i\| + \varepsilon \sup_{m\in\mathbb{N}} \|P_m - \mathbb{1}\| , \end{aligned} \tag{55}$$

where $\sup_{m\in\mathbb{N}} \|P_m - \mathbb{1}\| < \infty$ according to the theorem of Banach–Steinhaus. Furthermore, $\lim_{n\to\infty} \sup_{1\leq i\leq k} \|(P_n - \mathbb{1})y_i\| = 0$ holds by assumption, which implies that $\limsup_{n\to\infty} \sup_{y\in A} \|(P_n - \mathbb{1})y\| \leq \varepsilon \sup_{m\in\mathbb{N}} \|P_m - \mathbb{1}\|$, thus $\lim_{n\to\infty} \sup_{y\in A} \|(P_n - \mathbb{1})y\| = 0$, as ε was arbitrary. With this and (54) the claim follows. □

The question if such operators P_n exist—under the further restriction of being of finite rank—leads to

Definition 3. *If for a Banach space Y there exists a sequence (P_n) of operators of finite rank that satisfies the hypotheses of Proposition 3, then Y is said to have the* approximation property*, i.e., the operators of finite rank from X into Y are dense in the compact operators.*

This is not true for all Banach spaces, cf. the little review in [Wer00, Sec. II.6] or the advanced books by Lindenstrauss and Tzafriri [Lin77, Lin79]. An explicit proof for $\mathrm{L}^1(\mathbb{R})$ is given in Proposition 4 below. The more general case of L^p spaces, $1 \leq p \leq \infty$, is treated in [Sch74, Thm. IV.2.4].

2.2.2 Application to kernel operators

In our case of the COA model we have $X = Y = \mathrm{L}^1(\mathbb{R})$ and K is of the form

$$(Kf)(x) = \int_{\mathbb{R}} k(x, y) f(y) \, \mathrm{d}y \qquad \text{for all } x \in \mathbb{R} \tag{56}$$

with a measurable kernel $k\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$. Therefore, for the Galerkin method to work, it is necessary that $\mathrm{L}^1(\mathbb{R})$ has the approximation property. Generally, this is shown by means of conditional expectations in a quite abstract fashion, compare [Sch74, Thm. IV.2.4]. We will make this approach explicit here by using a sequence $(\{I_{n,k} : 1 \le k \le N_n\})_n$ of families of disjoint intervals that get finer and finer and at the same time ultimately cover every bounded interval.[7]

Proposition 4. *The Banach space* $\mathrm{L}^1(\mathbb{R})$ *has the approximation property, i.e., the operators of finite rank are dense in the compact operators in* $\mathrm{L}^1(\mathbb{R})$.

Explicitly, one may choose finite-dimensional subspaces Y_n *of* $Y = \mathrm{L}^1(\mathbb{R})$ *that consist of all step functions with prescribed (bounded) intervals* $I_{n,k}$ $(k \in \mathcal{N}_n := \{1, \dots, N_n\})$ *with the following properties:*

(I1) *For every bounded interval* $I \subset \mathbb{R}$ *and every* $\varepsilon > 0$ *there is an* n_0 *such that, for all* $n \ge n_0$, *a set* $L \subset \mathcal{N}_n$ *exists for which* $I_{n,L} := \bigcup_{\ell\in L} I_{n,\ell}$ *satisfies* $|I\backslash I_{n,L}| = 0$ *and* $|I_{n,L}\backslash I| < \varepsilon$. *(We then say that* I *is* ε-optimally covered.)

(I2) $|I_{n,k} \cap I_{n,\ell}| = 0$ *for all* $n \in \mathbb{N}$ *and* $1 \le k < \ell \le N_n$.

Then, with the characteristic functions $\varphi_{n,k} = 1_{I_{n,k}}$, *the projections* P_n *onto the subspaces* Y_n *spanned by* $\{\varphi_{n,k} : k \in \mathcal{N}_n\}$ *are given by*

$$P_n f = \sum_{k=1}^{N_n} \varphi_{n,k} \frac{1}{|I_{n,k}|} \int_{I_{n,k}} f(x)\,\mathrm{d}x \qquad \text{for } f \in \mathrm{L}^1(\mathbb{R}), \tag{57}$$

where $\int_{I_{n,k}} f(x)\,\mathrm{d}x$ *are the* conditional expectations *mentioned above. The projections satisfy* $\|P_n\|_1 = 1$.

Proof: Obviously, the subspaces Y_n are closed, finite-dimensional, and the P_n are, due to (I2), projections onto them. Since

$$\|P_n f\| = \sum_{k=1}^{N_n} \int_{I_{n,k}} |f(x)|\,\mathrm{d}x \le \int_{\mathbb{R}} |f(x)|\,\mathrm{d}x = \|f\|_1 \qquad \text{for every } f \in \mathrm{L}^1(\mathbb{R}) \tag{58}$$

and $\|P_n\varphi_{n,k}\| = \|\varphi_{n,k}\|$ for every $k \in \mathcal{N}_n$, we have $\|P_n\| = 1$.

We now show that $P_n \to \mathbb{1}$ pointwise. To this end, let $f \in \mathrm{L}^1(\mathbb{R})$ and $\varepsilon > 0$ be given. Remember that the set of all step functions is, by definition, dense in $\mathrm{L}^1(\mathbb{R})$, compare [Lan93, Sec. VI.3]. Therefore, we can find a step function $\psi = \sum_{k=1}^m \psi_k 1_{J_k}$ (with bounded intervals J_k) that satisfies $\|f - \psi\|_1 < \varepsilon/3$. Due to (I1) we can now choose an n_0 such that $|\bigcup_{k=1}^m J_k \backslash \bigcup_{k=1}^{N_n} I_{n,k}| = 0$ for all $n \ge n_0$. Then, the only contributions to $\|P_n\psi - \psi\|_1$ are due to mismatches at the boundaries of the J_k. Therefore, let J_k^+ and J_k^- $(k \in \mathcal{N}_n)$ be open intervals of measure $\eta = \varepsilon/(12m \max_k |\psi_k|)$ that contain the right and left boundary points of J_k, respectively. Choosing $n_1 \ge n_0$ according to (I1) large enough such that every $J_k^\pm$ is η-optimally covered for $n \ge n_1$, we have $\|P_n\psi - \psi\|_1 < 2\sum_{k=1}^m 2\eta\psi_k \le \varepsilon/3$ for $n \ge n_1$. Putting everything together yields, for $n \ge n_1$,

$$\|P_n f - f\|_1 \le \|P_n(f - \psi)\|_1 + \|P_n\psi - \psi\|_1 + \|\psi - f\|_1 < \varepsilon\,, \tag{59}$$

which proves $\|P_n f - f\| \to 0$ for $n \to \infty$ and thus the approximation property. □

[7] Both properties are formally captured by (I1) in Proposition 4.

With respect to a kernel operator K of the form (56) and some $f \in Y_n$, represented as $f = \sum_{k=1}^{N_n} \varphi_{n,k} f_k$, the above procedure amounts to the discretization

$$\begin{aligned}(P_n K f) &= \sum_{k=1}^{N_n} \varphi_{n,k} \frac{1}{|I_{n,k}|} \left(\int_{I_{n,k}} \int_{\mathbb{R}} k(x,y) \sum_{\ell=1}^{N_n} f_\ell \varphi_{n,\ell}(y) \, \mathrm{d}y \, \mathrm{d}x \right) \\ &= \sum_{k=1}^{N_n} \varphi_{n,k} \sum_{\ell=1}^{N_n} \frac{1}{|I_{n,k}|} \left(\int_{I_{n,k}} \int_{I_{n,\ell}} k(x,y) \, \mathrm{d}y \, \mathrm{d}x \right) f_\ell =: \sum_{k=1}^{N_n} \varphi_{n,k} \sum_{\ell=1}^{N_n} M_{n,k\ell} f_\ell \end{aligned} \tag{60}$$

with an $N_n \times N_n$ matrix $\boldsymbol{M}_n = (M_{n,k\ell})$. The corresponding eigenvalue equation is

$$\boldsymbol{M}_n \boldsymbol{g}_n = \nu_n \boldsymbol{g}_n \,, \qquad \text{or equivalently} \quad P_n K g_n = \nu_n g_n \,, \tag{61}$$

where $g_n \in Y_n$ is granted due to the projection property. An example of intervals $I_{n,k}$ satisfying (I1) and (I2) is $I_{n,k} = [-n + 2^{-n}(k-1), -n + 2^{-n}k]$ with $k \in \mathcal{N}_n = \{1, \ldots, 2^{n+1}n\}$.

2.2.3 Compact kernel operators

We further need to determine under which conditions a kernel operator K of the form (56) is compact on one of the following Banach spaces:

- $\mathrm{C}(\mathbb{R})$, equipped with the supremum norm $\|.\|_\infty$,
- $\mathrm{L}^1(\mathbb{R})$ with the usual norm $\|.\|_1$, or
- $\mathrm{C}_1(\mathbb{R}) := \mathrm{C}(\mathbb{R}) \cap \mathrm{L}^1(\mathbb{R})$ with the norm $|\!|\!|.|\!|\!| := \max\{\|.\|_\infty, \|.\|_1\}$.

To this end, we follow Jörgens [Jör70, Secs. 11, 12]. A first result in this respect is

Proposition 5 [Jör70, Thms. 12.2, 12.3]. *A kernel operator K on $\mathrm{C}(\mathbb{R})$ of the form (56) is compact if and only if the following two conditions are fulfilled:*

(C1) *The function $x \to k(x,.)$ from $\mathbb{R}$ to $\mathrm{L}^1(\mathbb{R})$ is continuous and bounded.*

(C2) *For every $\varepsilon > 0$ there exists a finite open covering $(V_1, \ldots, V_n)$ of $\mathbb{R}$ and points $x_j \in V_j$ such that $\|k(x,.) - k(x_j,.)\|_1 < \varepsilon$ for all $x \in V_j$ and all j.*

Then, its operator norm is given by

$$\|K\|_\infty = \sup_{x \in \mathbb{R}} \int_{\mathbb{R}} k(x,y) \, \mathrm{d}y \,. \tag{62}$$

As in [Jör70, Sec. 12.4], we consider the dual system $\langle \mathrm{C}(\mathbb{R}), \mathrm{C}_1(\mathbb{R}) \rangle$ with the bilinear form $\langle f, g \rangle = \int_{\mathbb{R}} f(x) g(x) \, \mathrm{d}x$. For this, we define the transposed K^{T} of K via

$$(K^{\mathrm{T}} g)(y) = \int_{\mathbb{R}} g(x) k(x,y) \, \mathrm{d}x \qquad \text{for all } y \in \mathbb{R}. \tag{63}$$

We then have

Theorem 5 (Jörgens). *If both K and K^{T} are compact as operators on $\mathrm{C}(\mathbb{R})$, they are so as operators on $\mathrm{C}_1(\mathbb{R})$ and $\mathrm{L}^1(\mathbb{R})$ as well. Then, they can be approximated by operators of finite rank.*

PROOF: As both K^{T} and K are bounded as operators on $\mathrm{C}(\mathbb{R})$, they are, at the same time, Hille–Tamarkin operators in $\mathcal{H}_{\infty\infty}(\mathbb{R})$ since the respective norm, $|\!|.|\!|_{\infty\infty}$ in (14), is just given by (62). Then, according to [Jör70, Thm. 11.5], K and K^{T} can also be regarded as bounded operators on $\mathrm{L}^1(\mathbb{R})$; thus, both map $\mathrm{C}_1(\mathbb{R})$ into itself. Due to [Jör70, Thm. 12.6] there is, for every $\varepsilon > 0$, an operator of finite rank, K_ε, with $|\!|\!|K_\varepsilon - K|\!|\!| < \varepsilon$, where $|\!|\!|A|\!|\!| := \max\{\|A\|_\infty, \|A^{\mathrm{T}}\|_\infty\}$ is a norm for the Banach algebra of all operators on $\mathrm{C}(\mathbb{R})$ that map $\mathrm{C}_1(\mathbb{R})$ into itself and have a transposed of the same kind. We have $|\!|\!|Af|\!|\!| \leq |\!|\!|A|\!|\!|\,|\!|\!|f|\!|\!|$ for $f \in \mathrm{C}_1(\mathbb{R})$, see [Jör70, Sec. 12.4]. Thus, $|\!|\!|A|\!|\!|$ can serve as an upper bound for the operator norm of A on $\mathrm{C}_1(\mathbb{R})$. Therefore, K is compact as an operator on $\mathrm{C}_1(\mathbb{R})$ and can be approximated by K_ε. Furthermore, according to [Jör70, Thm. 11.5], $\|K_\varepsilon - K\|_1 \leq |\!|\!|(K_\varepsilon - K)^{\mathrm{T}}|\!|\!| < \varepsilon$ holds. Hence, K is compact as an operator on $\mathrm{L}^1(\mathbb{R})$ as well. □

2.2.4 Application to the COA model

In order to be able to apply Theorem 5 to the COA model, we need both K_α and ${K_\alpha}^{\mathrm{T}}$ to be bounded as operators on $\mathrm{C}(\mathbb{R})$, i.e., cf. (62),

$$\|K_\alpha\|_\infty = \sup_{x\in\mathbb{R}} \int_{\mathbb{R}} k_\alpha(x,y)\,\mathrm{d}y < \infty\,, \tag{64}$$

$$\|{K_\alpha}^{\mathrm{T}}\|_\infty = \sup_{y\in\mathbb{R}} \int_{\mathbb{R}} k_\alpha(x,y)\,\mathrm{d}x < \infty\,. \tag{65}$$

The latter is already a consequence of (U4). Furthermore, both operators need to be compact, which can be checked by conditions (C1) and (C2).

With all this, we would be able to apply Theorem 4 and approximate K_α by operators of finite rank. However, the biological system is described by the (non-compact) operator $A = T - U$, not by some K_α. It will be shown that it is indeed possible to discretize the operators T and U directly by applying the projections P_n from Proposition 4, if further restrictions apply. Then, even more generally, the approximation can be done by choosing arbitrary points in the intervals $I_{n,k}$ at which the functions w and u are sampled. Both procedures will now be described in detail.

In the first setting, K_λ is approximated by $K_n := P_n U (P_n T + \lambda_n)^{-1}$. Explicitly, for $f \in Y_n$ with $f = \sum_{k=1}^{N_n} f_k \varphi_{n,k}$, it reads

$$P_n T f = \sum_{k=1}^{N_n} \varphi_{n,k} \frac{1}{|I_{n,k}|} \int_{I_{n,k}} w(x)\,\mathrm{d}x\, f_k = \sum_{k=1}^{N_n} \varphi_{n,k} w(t^w_{n,k}) f_k\,, \tag{66}$$

$$P_n U f = \sum_{k,\ell=1}^{N_n} \varphi_{n,k} \frac{1}{|I_{n,k}|} \int_{I_{n,k}} \int_{I_{n,\ell}} u(x,y)\,\mathrm{d}y\,\mathrm{d}x\, f_\ell = \sum_{k,\ell=1}^{N_n} \varphi_{n,k} |I_{n,\ell}| u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell}) f_\ell\,, \tag{67}$$

with appropriate points $t^w_{n,k}$, $t^{ux}_{n,k\ell} \in I_{n,k}$ and $t^{uy}_{n,k\ell} \in I_{n,\ell}$ that satisfy

$$\frac{1}{|I_{n,k}|} \int_{I_{n,k}} u(x, t^{uy}_{n,k\ell})\,\mathrm{d}x = u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell})\,. \tag{68}$$

These exist due to the continuity of w and u. But more generally, we may pick the points arbitrarily from the respective intervals.

In either case, we define the $N_n \times N_n$ matrices $\boldsymbol{T}_n$, $\boldsymbol{U}_n$, and $\boldsymbol{A}_n := \boldsymbol{T}_n - \boldsymbol{U}_n$ via

$$T_{n,kk} := w(t^w_{n,k})\,, \qquad U_{n,k\ell} := |I_{n,\ell}|\, u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell})\,. \tag{69}$$

The corresponding operators in Y_n are given by

$$T_n f = \sum_{k=1}^{N_n} \varphi_{n,k} T_{n,kk} f_k\,, \qquad U_n f = \sum_{k,\ell=1}^{N_n} \varphi_{n,k} U_{n,k\ell} f_\ell\,, \qquad A_n = T_n - U_n\,, \tag{70}$$

again with $f = \sum_{k=1}^{N_n} f_k \varphi_{n,k}$. For notational convenience, we also define the matrices $\boldsymbol{P}_{\alpha,n}$ by

$$P_n K_\alpha f = \sum_{k,\ell=1}^{N_n} \varphi_{n,k} \frac{1}{|I_{n,k}|} \int_{I_{n,k}} \int_{I_{n,\ell}} \frac{u(x,y)}{w(y)+\alpha}\,\mathrm{d}y\,\mathrm{d}x f_\ell =: \sum_{k,\ell=1}^{N_n} \varphi_{n,k} P_{\alpha,n,k\ell} f_\ell\,. \tag{71}$$

The eigenvalue equation to be solved is

$$(\boldsymbol{A}_n + \lambda_n)\boldsymbol{p}_n = 0\,, \tag{72}$$

which is equivalent to

$$(A_n + \lambda_n)p_n = 0\,, \tag{73}$$

where $p_n = \sum_{k=1}^{N_n} p_{n,k}\varphi_{n,k} \in Y_n$. With $K_{\alpha,n} = U_n(T_n+\alpha)^{-1}$, for $\alpha > -\min_{k\in\mathcal{N}_n} w(t_{n,k})$, and $q_n = (T_n + \lambda_n)p_n$ also the eigenvalue equation

$$(K_{\lambda_n,n} - 1)q_n = 0 \tag{74}$$

is solved (and vice versa), cf. Lemma 1. (The inequality $\lambda_n > -\min_{k\in\mathcal{N}_n} w(t_{n,k})$ follows from Theorem 1.)

For these procedures to be valid approximations, the first condition in (53), i.e., $\|K_n - P_n K\|_{Y_n} \to 0$, has to be true for $K = K_\lambda$ and $K_n = U_n(T_n+\lambda_n)^{-1}$. This, however, is not given automatically. Problems arise from the fact that in K_n the averaging defined by P_n (or, more generally, the sampling) is applied to the enumerator and denominator of k_{λ_n} separately, whereas in $P_n K$ the quotient k_λ is averaged as such. It turns out that some additional requirements of uniform continuity are sufficient for the convergence. This is made precise in the following two propositions.

Proposition 6. *Suppose that the following conditions are true:*

(S1) $u(x,.)$ *is uniformly continuous for all* $x \in \mathbb{R}$.
(S2) $u(.,y)$ *is continuous for all* $y \in \mathbb{R}$.
(S3) $(w+\alpha)^{-1}$ *is uniformly continuous for all* $\alpha > 0$.
(S4) *There is a function* $w_{\min}\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$, *satisfying*

$$\int_{\mathbb{R}} \sup_{y\in\mathbb{R}} \frac{u(x,y)}{w_{\min}(y)+\alpha}\,\mathrm{d}x < \infty \qquad \textit{for all } \alpha > 0, \tag{75}$$

and an $n_0 \in \mathbb{N}$ *such that* $w(y) \geq w_{\min}(y')$ *for all* $n \geq n_0$, $\ell \in \mathcal{N}_n$, *and* $y, y' \in I_{n,\ell}$.

Then, for $K = K_\alpha$ and $K_n = P_n U(P_n T + \alpha)^{-1}$ with any $\alpha > 0$ and the projections P_n from Proposition 4, the first condition in (53) *is fulfilled, i.e., $\|K_n - P_n K\|_{Y_n} \to 0$ as $n \to \infty$. The same is true for $K_n = K_{\alpha,n} = U_n(T_n + \alpha)^{-1}$ with the more general discretization from above if in addition to* (S1)–(S4) *the following condition is satisfied:*

(S5) *There is a function $u_{\max} \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, satisfying*

$$\int_{\mathbb{R}} \sup_{y \in \mathbb{R}} \frac{u_{\max}(x,y)}{w_{\min}(y) + \alpha} \,\mathrm{d}x < \infty \qquad \text{for all } \alpha > 0, \tag{76}$$

and an $n_1 \geq n_0$ such that $u(x,y) \leq u_{\max}(x',y)$ for all $n \geq n_1$, $k \in \mathcal{N}_n$, $y \in \mathbb{R}$, and x, $x' \in I_{n,k}$.

Let us split the rather technical proof into a couple of digestible lemmas.

Lemma 8. *If conditions* (S1)–(S3) *are true, then for every $\varepsilon > 0$ and every compact interval $I \subset \mathbb{R}$ there is an n_2 such that for all $n \geq n_2$ and all k, $\ell \in \mathcal{N}_n$ with $I_{n,k} \cap I \neq \emptyset$ we have*

$$\frac{1}{|I_{n,\ell}|} \left| P_{\alpha,n,k\ell} - \frac{U_{n,k\ell}}{T_{n,\ell\ell} + \alpha} \right| < \frac{\varepsilon}{|I|}\,. \tag{77}$$

Proof: Let ε and I be given as above and $I_0 = \bigcup_{n \in \mathbb{N}} \bigcup_{k : I_{n,k} \cap I \neq \emptyset} I_{n,k}$, which is a bounded interval due to (I1) from Proposition 4. By assumptions (S1)–(S3), u and $(w + \alpha)^{-1}$ are uniformly continuous on $\overline{I_0} \times \mathbb{R}$ and $\mathbb{R}$, respectively. Further, $(w + \alpha)^{-1}$ is bounded by α^{-1}. Thus, there is an n_2 such that, for every $n \geq n_2$ and k, $\ell \in \mathcal{N}_n$ with $I_{n,k} \cap I \neq \emptyset$,

$$\begin{aligned} &\left| \frac{1}{|I_{n,k}|} \frac{1}{|I_{n,\ell}|} \int_{I_{n,k}} \int_{I_{n,\ell}} \frac{u(x,y)}{w(y) + \alpha} \,\mathrm{d}y \,\mathrm{d}x - \frac{u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell})}{w(t^{w}_{n,\ell}) + \alpha} \right| \\ &\qquad = \left| \frac{u(t^{kx}_{n,k\ell}, t^{ky}_{n,k\ell})}{w(t^{ky}_{n,k\ell}) + \alpha} - \frac{u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell})}{w(t^{w}_{n,\ell}) + \alpha} \right| \\ &\qquad \leq \left| \frac{u(t^{kx}_{n,k\ell}, t^{ky}_{n,k\ell})}{w(t^{ky}_{n,k\ell}) + \alpha} - \frac{u(t^{ux}_{n,k\ell}, t^{w}_{n,\ell})}{w(t^{w}_{n,\ell}) + \alpha} \right| + \frac{|u(t^{ux}_{n,k\ell}, t^{w}_{n,\ell}) - u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell})|}{w(t^{w}_{n,\ell}) + \alpha} < \frac{\varepsilon}{|I|}\,. \end{aligned} \tag{78}$$

Here, the points $t^{kx}_{n,k\ell} \in I_{n,k}$ and $t^{ky}_{n,k\ell} \in I_{n,\ell}$ are chosen such that the first equality holds, which is possible due to the continuity of k_α. From this the claim follows easily with (69) and (71). □

Lemma 9. *For every $\varepsilon > 0$ there is a compact interval I_1 such that, for all intervals $I \supset I_1$ and all $n \in \mathbb{N}$,*

$$\sum_{\substack{k \\ I_{n,k} \cap I = \emptyset}} |I_{n,k}| \max_{\ell \in \mathcal{N}_n} \frac{P_{\alpha,n,k\ell}}{|I_{n,\ell}|} < \varepsilon\,. \tag{79}$$

Proof: Due to (U4) there is a compact interval I_1 such that, for all $I \supset I_1$,

$$\begin{aligned} \sum_{\substack{k \\ I_{n,k} \cap I = \emptyset}} |I_{n,k}| \max_{\ell \in \mathcal{N}_n} \frac{P_{\alpha,n,k\ell}}{|I_{n,\ell}|} &\leq \sum_{\substack{k \\ I_{n,k} \cap I = \emptyset}} |I_{n,k}| \frac{1}{|I_{n,k}|} \int_{I_{n,k}} \max_{y \in \mathbb{R}} \frac{u(x,y)}{w(y) + \alpha} \,\mathrm{d}x \\ &\leq \int_{\mathbb{R} \setminus I_1} \max_{y \in \mathbb{R}} \frac{u(x,y)}{w(y) + \alpha} \,\mathrm{d}x < \varepsilon\,. \end{aligned} \tag{80}$$

□

Lemma 10. *If condition* (S4) *is true, and if*
(i) $U_{n,k\ell} = \frac{|I_{n,\ell}|}{|I_{n,k}|}\int_{I_{n,k}} u(x, t^{uy}_{n,k\ell})\,\mathrm{d}x$ *for all* k, $\ell \in \mathcal{N}_n$ *or*
(ii) *condition* (S5) *is fulfilled,*
then for every $\varepsilon > 0$ *there is a compact interval* I_2 *such that, for all intervals* $I \supset I_2$ *and all* $n \in \mathbb{N}$,

$$\sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} |I_{n,k}| \max_{\ell\in\mathcal{N}_n} \frac{U_{n,k\ell}}{|I_{n,\ell}|(T_{n,\ell\ell}+\alpha)} < \varepsilon\,. \tag{81}$$

Proof: In case (i) we have, using (68),

$$\begin{aligned}\sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} |I_{n,k}| \max_{\ell\in\mathcal{N}_n} \frac{u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell})}{w(t^w_{n,\ell})+\alpha} &= \sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} \max_{\ell\in\mathcal{N}_n} \frac{\int_{I_{n,k}} u(x, t^{uy}_{n,k\ell})\,\mathrm{d}x}{w(t^w_{n,\ell})+\alpha}\\ \le \sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} \int_{I_{n,k}} \max_{y\in\mathbb{R}} \frac{u(x,y)}{w_{\min}(y)+\alpha}\,\mathrm{d}x &\le \int_{\mathbb{R}\setminus I_2} \max_{y\in\mathbb{R}} \frac{u(x,y)}{w_{\min}(y)+\alpha}\,\mathrm{d}x < \varepsilon\end{aligned} \tag{82}$$

for some compact interval I_2, due to (S4), and all intervals $I \supset I_2$. In case (ii) we can find, due to (S5), a compact interval I_2 such that, for all intervals $I \supset I_2$,

$$\begin{aligned}\sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} |I_{n,k}| \max_{\ell\in\mathcal{N}_n} \frac{u(t^{ux}_{n,k\ell}, t^{uy}_{n,k\ell})}{w(t^w_{n,\ell})+\alpha} &\le \sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} |I_{n,k}| \max_{y\in\mathbb{R}} \frac{u(t^{ux}_{n,k\ell}, y)}{w_{\min}(y)+\alpha}\\ \le \sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} \max_{y\in\mathbb{R}} \frac{\int_{I_{n,k}} u_{\max}(x,y)\,\mathrm{d}x}{w_{\min}(y)+\alpha} &\le \int_{\mathbb{R}\setminus I_2} \max_{y\in\mathbb{R}} \frac{u_{\max}(x,y)}{w_{\min}(y)+\alpha}\,\mathrm{d}x < \varepsilon\,.\end{aligned} \tag{83}$$

□

Proof of Proposition 6: Let $\varepsilon > 0$ be given. Choose a compact interval I such that $I \supset I_1 \cup I_2$ with I_1 and I_2 from Lemmas 9 and 10, respectively. Let $I_3 = \overline{\bigcup_{n,k:I_{n,k}\cap I\neq\emptyset} I_{n,k}}$. Further, let n_0 be as in (S4), n_1 as in (S5) (or $n_0 = n_1$ if not applicable), n_2 as in Lemma 8, and $n \ge \max\{n_0, n_1, n_2\}$. Then

$$\begin{aligned}\|P_n K_\alpha - K_{\alpha,n}\|_{Y_n} &= \sup_{\substack{f\in Y_n\\ \|f\|_{Y_n}\le 1}} \sum_{k=1}^{N_n} |I_{n,k}| \left| \sum_{\ell=1}^{N_n} \left(P_{\alpha,n,k\ell} - \frac{U_{n,k\ell}}{T_{n,\ell\ell}+\alpha}\right) f_\ell \right|\\ &\le \left(\sum_{\substack{k\\ I_{n,k}\cap I\neq\emptyset}} + \sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}}\right) |I_{n,k}| \max_{\ell\in\mathcal{N}_n} \frac{1}{|I_{n,\ell}|} \left| P_{\alpha,n,k\ell} - \frac{U_{n,k\ell}}{T_{n,\ell\ell}+\alpha} \right|\\ &\le \sum_{\substack{k\\ I_{n,k}\cap I\neq\emptyset}} |I_{n,k}| \frac{\varepsilon}{|I_3|} + \sum_{\substack{k\\ I_{n,k}\cap I=\emptyset}} |I_{n,k}| \left(\max_{\ell\in\mathcal{N}_n} \frac{P_{\alpha,n,k\ell}}{|I_{n,\ell}|} + \max_{\ell\in\mathcal{N}_n} \frac{U_{n,k\ell}}{|I_{n,\ell}|(T_{n,\ell\ell}+\alpha)} \right) < 3\varepsilon\end{aligned} \tag{84}$$

according to Lemmas 8–10. From this the claim follows. □

Proposition 7. *Let* $\alpha_n > -\min_{k\in\mathcal{N}_n} w(t_{n,k})$ *with* $\alpha_n \to \alpha > 0$ *as* $n \to \infty$ *and the hypotheses of Proposition* 6 *be satisfied. Then* $\|K_{\alpha_n,n} - P_n K_\alpha\|_{Y_n} \to 0$.

PROOF: Consider

$$\begin{aligned}\|P_nK_\alpha - &U_n(T_n+\alpha_n)^{-1}\|_{Y_n}\\ &\le \|P_nK_\alpha - U_n(T_n+\alpha)^{-1}\|_{Y_n} + \|U_n[(T_n+\alpha_n)^{-1} - (T_n+\alpha)^{-1}]\|_{Y_n}\,.\end{aligned} \tag{85}$$

The first term tends to zero as $n \to \infty$ according to Proposition 6. For the second, choose n_0 such that $\inf_{n\ge n_0}\alpha_n > 0$. Then, for $n \ge n_0$,

$$\begin{aligned}\|U_n[(T_n+\alpha_n)^{-1} &- (T_n+\alpha)^{-1}]\|_{Y_n}\\ &= |\alpha-\alpha_n|\,\|U_n(T_n+\alpha_n)^{-1}(T_n+\alpha)^{-1}\|_{Y_n} \le |\alpha-\alpha_n|\,\|U\|_Y(\inf_{n\ge n_0}\alpha_n)^{-1}\alpha^{-1}\,.\end{aligned} \tag{86}$$

This vanishes as $n \to \infty$ since all constants that occur are finite, from which the claim follows. □

2.2.5 Convergence of eigenvalues and eigenvectors

Let us now show

Theorem 6. *Let* λ, p, λ_n, *and* p_n *be as in* (7) *and* (73)*. Then*

(a) $\lim_{n\to\infty}\lambda_n = \lambda > 0$ *and*
(b) $\lim_{n\to\infty}\|p_n - p\|_1 = 0$, *i.e., the probability measures corresponding to these densities converge in total variation.*

The plan is the same as described in Section 2.1.3. The proofs, however, are quite different due to the more general setup.

Lemma 11. *There is a constant* $M > 0$ *such that* $\limsup_{n\to\infty}\lambda_n \le M$.

PROOF: Choose an $\alpha > 0$ such that $\|K_\alpha\|_Y \le 1-\varepsilon$ for some $0<\varepsilon<1$, which is possible since $\|K_\alpha\|_Y \to 0$ for $\alpha\to\infty$. Then, due to Propositions 3 and 6, respectively, there is an n_0 such that $|\|P_nK_\alpha\|_Y - \|K_\alpha\|_Y| \le \varepsilon/3$ and $|\|K_{\alpha,n}\|_{Y_n} - \|P_nK_\alpha\|_{Y_n}| \le \varepsilon/3$ for all $n \ge n_0$. For these n we have $\rho(K_{\alpha,n}) \le \|K_{\alpha,n}\|_{Y_n} \le \|P_nK_\alpha\|_{Y_n} + \varepsilon/3 \le \|P_nK_\alpha\|_Y + \varepsilon/3 \le \|K_\alpha\|_Y + 2\varepsilon/3 \le 1 - \varepsilon/3 < 1$ and thus $\lambda_n < \alpha$ by Lemma 2. Then, with $M = \alpha$, the claim follows. □

Lemma 12. $\liminf_{n\to\infty}\lambda_n > 0$.

PROOF: Similarly as in the proof of Lemma 5, we choose $\alpha > 0$ such that $\rho(K_\alpha) \ge 1+\varepsilon$ with a sufficiently small $\varepsilon > 0$. We know from the theorem of Jentzsch [Sch74, Thm. V.6.6] that $\rho(K_\alpha)$ is a simple eigenvalue of K_α and the only one with a positive eigenfunction. The same is true for $\rho(K_{\alpha,n})$ with respect to $K_{\alpha,n}$ (as an operator in Y_n). Theorem 4 together with Proposition 6 implies that there is a sequence of eigenvalues ν_n of $K_{\alpha,n}$ with limit $\rho(K_\alpha)$. Therefore, $\liminf_{n\to\infty}\rho(K_{\alpha,n}) \ge \rho(K_\alpha) \ge 1+\varepsilon$ and thus $\lambda_n > \alpha > 0$ for sufficiently large n. From this the claim follows. □

Proof of Theorem 6(a): From Lemma 11 and 12 we conclude that there is a convergent subsequence $(\lambda_{n_i})_i$ with limit $\lambda' \in]0, M]$. Then, due to Proposition 7, $K_{\lambda_{n_i},n_i}$ converges to $P_n K_{\lambda'}$ in norm. Hence, with Theorem 4, $\lim_{i\to\infty} \rho(K_{\lambda_{n_i},n_i}) = \rho(K_{\lambda_{n_i},n_i}) = 1$ is an eigenvalue of $K_{\lambda'}$. Furthermore, a subsequence of q_{n_i} converges to an eigenfunction $\tilde{q}$ of $K_{\lambda'}$, and $\tilde{q} \geq 0$ (but $\tilde{q} \neq 0$). As there is only one non-negative eigenfunction, we conclude $\lambda' = \lambda$. Since this is true for every convergent subsequence of (λ_n), the claim follows. □

Lemma 13. *There are constants c_1, $c_2 > 0$ and an $n_0 \in \mathbb{N}$, such that, for every $n \geq n_0$, $0 < c_1\|p_n\|_1 \leq \|q_n\|_1 \leq c_2\|p_n\|_1$ holds.*

Proof: Choose n_0 according to Lemma 12 such that $c_1 := \inf_{n\geq n_0} \lambda_n > 0$. One then easily checks that, for $n \geq n_0$,

$$\begin{aligned} 0 < c_1\|p_n\|_1 = c_1\|(P_nT + \lambda_n)^{-1}q_n\|_1 &\leq \|q_n\|_1 \\ = \|(P_nT + \lambda_n)p_n\|_1 = \|P_nUp_n\|_1 &\leq \|U\|_Y\|p_n\|_1 =: c_2\|p_n\|_1 \,. \end{aligned} \tag{87}$$

□

Lemma 14. *If $\|q_n\|_1 \to c$ holds, then also $q_n \to c\, q/\|q\|_1$.*

Proof: Due to the compactness of K_λ, there is a subsequence $(q_{n_i})_i$ such that $(K_\lambda q_{n_i})_i$ converges. Let $\tilde{q}$ denote its limit. Then

$$\begin{aligned} \|q_{n_i} - \tilde{q}\|_1 &\leq \|K_{n_i}q_{n_i} - P_{n_i}K_\lambda q_{n_i}\|_1 + \|(P_{n_i}K_\lambda - K_\lambda)q_{n_i}\|_1 + \|K_\lambda q_{n_i} - \tilde{q}\|_1 \\ &\leq \|K_{n_i} - P_{n_i}K_\lambda\|_{Y_{n_i}}\|q_{n_i}\|_1 + \|P_{n_i}K_\lambda - K_\lambda\|_Y\|q_{n_i}\|_1 + \|K_\lambda q_{n_i} - \tilde{q}\|_1 \to 0 \end{aligned} \tag{88}$$

due to Propositions 7 and 4. Therefore, we have $\lim_{i\to\infty} q_{n_i} = \tilde{q}$, with $\|\tilde{q}\|_1 = c$ and $\tilde{q} \geq 0$ a.e. Now we continue as in the proof of Lemma 7. Consider

$$\begin{aligned} &\|\tilde{q} - K_\lambda\tilde{q}\|_1 \\ &\leq \|\tilde{q} - K_{n_i}q_{n_i}\|_1 + \|K_{n_i} - P_{n_i}K_\lambda\|_{Y_n}\|q_{n_i}\|_1 + \|K_\lambda\|_Y\|q_{n_i} - \tilde{q}\|_1 + \|P_{n_i}K_\lambda - K_\lambda\|_Y\|\tilde{q}\|_1 \,. \end{aligned}$$

The first term vanishes for $i \to \infty$ due to $K_{n_i}q_{n_i} = q_{n_i}$, the second and fourth according to Propositions 7 and 4, respectively. Thus, $\tilde{q}$ is an eigenfunction of K_λ to the eigenvalue 1 and, according to Theorem 1, unique up to normalization, and hence $\tilde{q} = q$. Since this is true for all convergent subsequences, the claim follows (again due to the compactness of K_λ). □

Proof of Theorem 6(b): With our choice $\|p_n\|_1 = 1$ there is, due to Lemma 13, a subsequence $(q_{n_i})_i$ such that $\|q_{n_i}\|_1$ converges and has a strictly positive limit c. Thus, according to Lemma 14, also $q_{n_i} \to cq/\|q\|_1 =: \tilde{q}$ holds. Let n_0 be sufficiently large such that $\alpha := \inf_{n\geq n_0} \lambda_n > 0$. Then, for $n \geq n_0$,

$$\begin{aligned} \|p_n - (T+\lambda)^{-1}\tilde{q}\|_1 &= \|(P_nT + \lambda_n)^{-1}q_n - (T+\lambda)^{-1}\tilde{q}\|_1 \\ &\leq \|[(P_nT+\lambda_n)^{-1} - (P_nT+\lambda)^{-1}]q_n\|_1 \\ &\quad + \|(P_nT+\lambda)^{-1}(q_n - \tilde{q})\|_1 \\ &\quad + \|[(P_nT+\lambda)^{-1} - (T+\lambda)^{-1}]\tilde{q}\|_1 \\ &\leq \tfrac{1}{\alpha\lambda}|\lambda - \lambda_n|c_2 + \tfrac{1}{\lambda}\|q_n - \tilde{q}\|_1 + \tfrac{1}{\lambda^2}\|(\mathbb{1} - P_n)T\tilde{q}\|_1 \to 0 \,. \end{aligned} \tag{89}$$

With this, we have $p_{n_i} \to (T+\lambda)^{-1}\tilde{q}$ in $\mathrm{L}^1(I)$, and especially $\|(T+\lambda)^{-1}\tilde{q}\|_1 = 1$, from which $c = \|q\|_1$ follows. Therefore, each convergent subsequence of $(q_n)_n$ converges to q and thus, as in the proof of Lemma 14, the sequence itself. Together with what has just been shown the claim follows. □

2.2.6 Comparison to the case of a compact genotype interval

Both approaches, the application of the Nyström method in the case of a compact genotype interval and of the Galerkin method in the case of an unbounded interval, effectively lead to the same approximation procedure in our case of the COA model. First, one chooses appropriate intervals $I_{n,k}$ and points $t_{n,k} \in I_{n,k}$ (also for an unbounded interval the use of identical points $t^w_{n,k} = t^{ux}_{n,k\ell} = t^{uy}_{n,\ell k} = t_{n,k}$ seems reasonable in many cases). Then, the operators T and U from (4) and (5), respectively, are approximated by matrices $\boldsymbol{T}_n$ and $\boldsymbol{U}_n$, cf. (30), (31), and (69). For these, the eigenvalue problem $(\boldsymbol{T}_n - \boldsymbol{U}_n + \lambda_n)\boldsymbol{p}_n = 0$ is solved. Here, the eigenvectors $\boldsymbol{p}_n$ are considered as probability *densities* on I. Then, under the conditions described above, the eigenvalues λ_n converge to λ and the measures corresponding to the $\boldsymbol{p}_n$ converge in total variation to the equilibrium genotype distribution described by the solution p of the original problem (2).

The differences between the two approaches lie on the intermediate technical level of the compact operators K_α and $K_{\alpha,n}$ and the solutions q and q_n of the equivalent eigenvalue problems (13), (27), and (52). Here, in the first case we have collectively compact convergence $K_{\lambda_n,n} \xrightarrow{\mathrm{cc}} K_\lambda$ going together with $\|q_n - q\|_\infty \to 0$, whereas in the second case $\|P_n K_\lambda - K_\lambda\|_Y \to 0$ in $Y = \mathrm{L}^1(\mathbb{R})$ and $\|K_{\lambda_n,n} - P_n K_\lambda\|_{Y_n} \to 0$ in the subspaces Y_n going together with $\|q_n - q\|_1 \to 0$. On this level, neither does $\|K_{\lambda_n,n} - K_\lambda\|_\infty \to 0$ hold in the first case, compare [Kre99, Thm. 12.8], nor any kind of collectively compact convergence in the second.

Both methods may, strictly speaking, only be applied to continuous mutation kernels u. This excludes, for example, Γ-distributions (reflected at the source type), where $u(x,y) \propto |x-y|^{\Theta-1}\exp(-d\,|x-y|)$, which have poles for $x = y$ if $\Theta \in \,]0,1[$ and $d > 0$. These distributions incorporate biologically desirable properties, such as strong leptokurticity, and have been used, e.g., in [Hil82]. However, kernels as the above may be approximated arbitrarily well by continuous ones in the sense that the norm of the difference operator—and thus the difference of the largest eigenvalues—gets arbitrarily small. Then, the procedures described here may be applied to these continuous kernels.

3 Towards a simple maximum principle

In Section I.3.4 a simple maximum principle was derived that characterizes the equilibrium mean fitness of the mutation–selection model with discrete genotypes. With the results of the previous two sections, one can hope to transfer it, at least partially, to the COA model. Some first results in this direction are presented in the rest of this section. In Section 3.1, an upper bound for the mean fitness is given, which is valid if the mutation kernel may be symmetrized by a multiplication operator. Then, a symmetric analog of the equilibrium condition (2) exists, which involves the ancestral distribution. Section 3.2 is concerned with lower bounds. Here, weaker restrictions are sufficient, namely some kind of approximate local symmetrizability of the mutation kernel. Then, however, the interpretation of the symmetrized system in terms of the ancestral distribution is lost and—so far—no sharp upper bound can be proved.

In the mutation class limit of the discrete case, upper and lower bounds could be shown to converge towards each other, which established the maximum principle. Therefore, a natural question is whether there is an analogous limit for the COA model. So far, the situation is clear only in the very restricted setting that the mutant distributions—relative to the source type—are translationally invariant. The positive answer for this case is discussed in Section 3.3. The numerical examples in Section 3.4, however, lead to the conjecture that local symmetrizability is sufficient for the existence of a simple maximum principle.

In this section we come back to the notational convention from Chapter I and use the (non-compact) operator $H = U - T$, with T and U from (4) and (5), instead of $A = -H$ from (6).

3.1 An upper bound for the mean fitness

In the discrete case, the derivation of an upper bound for the mean fitness was based on the symmetric equilibrium condition (I.30) around a local maximum of the ancestral distribution. A similar argument is possible for the COA model—both with a compact and with an unbounded genotype interval. As for the discrete model, the approach relies on the possibility to symmetrize the operator H by means of a multiplication operator. This is analyzed in the following

Proposition 8. *Assume the notation from Section 1.1 and suppose the conditions from Section 1.2 are true. Then there is a multiplication operator S, defined via a function $s \in \mathrm{C}(I)$ by $(Sf)(x) = \exp(s(x))f(x)$, such that $\tilde{H} := SHS^{-1}$ is symmetric if and only if the mutation kernel is of the form*

$$u(x,y) = \exp(\beta(x))\, \tilde{u}(x,y)\, \exp(-\beta(y)) \tag{90}$$

with $\beta \in \mathrm{C}(I)$ and a continuous, symmetric function $\tilde{u}\colon I \times I \to \mathbb{R}_{\geq 0}$, $\tilde{u}(x,y) = \tilde{u}(y,x)$. In this case, $s = -\beta + c$ with some $c \in \mathbb{R}$ and $\tilde{u}$ is uniquely determined by

$$\tilde{u}(x,y) = \sqrt{u(x,y)u(y,x)}\,. \tag{91}$$

For the proof we need the following

Lemma 15. *Any function $u\colon I \times I \to \mathbb{R}_{\geq 0}$ with symmetric zeros, i.e., for which*

$$u(x,y) = 0 \quad \Leftrightarrow \quad u(y,x) = 0\,, \tag{92}$$

can be written in the form

$$u(x,y) = \tilde{u}(x,y)\exp(\alpha(x,y)) \tag{93}$$

with $\tilde{u}\colon I \times I \to \mathbb{R}_{\geq 0}$ being symmetric, $\tilde{u}(y,x) = \tilde{u}(x,y)$, and $\alpha\colon I \times I \to \mathbb{R}$ being anti-symmetric, $\alpha(y,x) = -\alpha(x,y)$. The function $\tilde{u}$ is uniquely determined by the condition $\tilde{u}(x,y) = \sqrt{u(x,y)u(y,x)}$ and $\alpha(x,y)$ is only ambiguous for points where $u(x,y) = 0$. If u is continuous, $\tilde{u}$ is continuous as well and α is continuous at least where $u > 0$.

PROOF: Let $\tilde{u}(x,y) = \sqrt{u(x,y)u(y,x)}$, which is continuous if u is. If $\tilde{u}(x,y) > 0$ let $\alpha(x,y) = \frac{1}{2}(\log u(x,y) - \log u(y,x))$. Then (93) holds as $\exp(\alpha(x,y)) = \sqrt{u(x,y)/u(y,x)}$, and α is continuous in a neighborhood of (x,y) if u is. Otherwise, (93) is trivially true due to (92), and $\alpha(x,y)$ is arbitrary (and can thus be chosen anti-symmetric). Suppose there are functions $\hat{u}$ and β with the same properties as $\tilde{u}$ and α. Then these satisfy $\tilde{u}(x,y) = \sqrt{\hat{u}(x,y)\exp(\beta(x,y))\hat{u}(y,x)\exp(\beta(y,x))} = \hat{u}(x,y)$ and, if $\tilde{u}(x,y) > 0$, also $2\alpha(x,y) = \log\hat{u}(x,y) + \beta(x,y) - \log\hat{u}(y,x) - \beta(y,x) = 2\beta(x,y)$. □

PROOF OF PROPOSITION 8: Note first that, if s exists, it is required to be bounded and thus both S and S^{-1} are bounded as operators in $\mathrm{C}(I)$ or $\mathrm{L}^1(I)$. Further, the operator $\tilde{H}$ is symmetric if and only if $\tilde{U} = SUS^{-1}$ is. With the notation from Lemma 15, which is applicable since the symmetry of $\tilde{U}$ implies that u has symmetric zeros, this is the case if and only if

$$\exp(s(x))\,\tilde{u}(x,y)\exp(\alpha(x,y))\exp(-s(y)) = \exp(s(y))\,\tilde{u}(y,x)\exp(\alpha(y,x))\exp(-s(x)) \tag{94}$$

for all x, $y \in I$, with $\tilde{u}$ satisfying (91). For $\tilde{u}(x,y) > 0$, this is equivalent to

$$\alpha(x,y) = \beta(x) - \beta(y) \tag{95}$$

with $\beta = -s - c$ and any $c \in \mathbb{R}$. Assume s exists as described. Then we have the freedom to choose α according to (95) even for $\tilde{u}(x,y) = 0$. Conversely, if (90) is true, any $s = -\beta + c$ has the desired properties. □

If there is a symmetric $\tilde{H}$, the equivalent of the equilibrium condition (2) is

$$\big(r(x) - u_1(x)\big)\sqrt{a(x)} + \int_I \sqrt{u(x,y)u(y,x)}\sqrt{a(y)}\,\mathrm{d}y = \lambda\sqrt{a(x)} \qquad \text{for all } x \in I\,, \tag{96}$$

which follows from multiplying (2) by $\exp(-\beta(x))$ and letting $\sqrt{a(x)} = \exp(-\beta(x))p(x)$. Here, the additive constant in β should be chosen such that a is a probability density, i.e., $\int_I a(x)\,\mathrm{d}x = \int_I(\exp(-\beta(x))p(x))^2\,\mathrm{d}x = 1$. Indeed, a describes the ancestral distribution with the same interpretation as in Section I.2.3. Note that even in the case of an unbounded genotype interval, if conditions (S1)–(S3) on page 49 are true, Proposition 1 can easily be generalized, which implies that p, and thus a, is continuous. Now, with (96), an upper bound for λ can be established.

Theorem 7. *Suppose there is an operator S as specified in Proposition 8 and, if the genotype interval I is unbounded, that the ancestral density a is bounded and assumes its supremum. Then an upper bound for the mean fitness λ is given by*

$$\lambda \le \sup_{x\in I}\big(r(x) - g(x)\big) \tag{97}$$

with

$$g(x) = \int_I u(y,x)\,\mathrm{d}y - \int_I \sqrt{u(x,y)u(y,x)}\,\mathrm{d}y\,. \tag{98}$$

PROOF: Choose $\hat{x}$ such that $a(\hat{x}) = \sup_{x\in I} a(x)$. Then, (96) with $x = \hat{x}$ yields

$$\lambda \le \big(r(\hat{x}) - u_1(\hat{x})\big) + \int_I \tilde{u}(\hat{x},y)\,\mathrm{d}y = r(\hat{x}) - g(\hat{x}) \le \sup_{x\in I}\big(r(x) - g(x)\big)\,. \tag{99}$$

□

3.2 A lower bound for the mean fitness

In this section, we restrict ourselves to a compact genotype interval I and continuous functions r and u. Furthermore, we need the mutation kernel to be symmetrizable—either globally, which amounts to (90), or locally in some sense, as analyzed in Proposition 9 below. The latter requires u to be of the form

$$u(x,y) = \exp\bigl(\gamma(y)(x-y)\bigr)h(y,|x-y|) \tag{100}$$

or

$$u(x,y) = \exp\Bigl(\gamma\bigl(\tfrac{1}{2}(x+y)\bigr)\tfrac{1}{2}(x-y)\Bigr)\tilde{u}(x,y)\,, \tag{101}$$

where $\tilde{u}(x,y) = \sqrt{u(x,y)u(y,x)}$ is symmetric.

In (100), each function $h(y,.)\colon [0,|I|] \to \mathbb{R}_{\geq 0}$ describes the common shape of the mutant distribution to the left and right of the source genotype y. Any asymmetry between both sides may only be due to an exponential in the difference of the destination genotype x and y, where the exponential factor γ may depend on y. Similarly, in (101), $\tilde{u}$ describes the symmetric part when interchanging x and y, and γ, here depending on the arithmetic mean of x and y, may lead to some asymmetry. Although (100) and (101) are quite strong restrictions, they allow for many biologically reasonable mutation kernels, going beyond symmetric Gaussian and Γ-distributions (reflected about zero), which are usually considered, cf. [Bür00, Fig. IV.2.1]. An example, also displaying different degrees of asymmetry, is shown in the right panel of Figure 2 in Section 3.4 below.

The main result concerning lower bounds for the mean fitness is

Theorem 8. *Consider the COA model as described by* (2) *with a compact genotype interval I and continuous functions r and u. Suppose the mutation kernel is of the form* (90), (100), *or* (101). *Let $\varepsilon > 0$ be given. Then, lower bounds for the mean fitness λ are given by*

$$\lambda \geq r(z) - g_\varepsilon(z) \tag{102}$$

with z such that $J_z := [z-\eta/2, z+\eta/2] \subset I$ and

$$g_\varepsilon(z) = u_1(z) - 2\int_0^{\frac{\eta}{2}} h(z,\xi)\,\mathrm{d}\xi + 2\int_0^{\frac{\eta}{2}}\int_\xi^{\frac{\eta}{2}} |h(z,\zeta) - h(z,\eta-\zeta)|\,\mathrm{d}\zeta\,\mathrm{d}\xi + \varepsilon\,. \tag{103}$$

Here, the largest possible value of $\eta > 0$ depends on ε, r, and u and, in the cases (90) *and* (101),

$$h(z,\xi) = \tilde{u}(z+\xi/2, z-\xi/2)\,. \tag{104}$$

The strategy for the proof of Theorem 8 is based on the way we proceeded in the discrete case (Section I.3.4). We look at small genotype intervals, for which the largest eigenvalues of the corresponding local subsystems serve as lower bounds for λ. Their common length η is chosen small enough such that the reproduction rates and mutant distributions can be considered constant—up to a part that contributes less than a given ε to the largest eigenvalue of the subsystem. For these (approximate) subsystems, a lower bound for the largest eigenvalue can be given explicitly.

First, we need some definitions. Let $w_0 = \max_{x\in I} w(x)$, then $H + w_0$ is a positive operator. Furthermore, it is convenient to consider $\mathrm{C}(I)$ as a subspace of $\mathrm{L}^\infty(I)$, since then, for a compact subinterval $J_z = [z-\eta/2, z+\eta/2] \subset I$, we can define the projection onto the subspace $\mathrm{L}^\infty(J_z)$ as $P_z\colon \mathrm{L}^\infty(I) \to \mathrm{L}^\infty(I)$, $f \mapsto 1_{J_z} f$ and $A_z = P_z A P_z$ as the corresponding restriction of an operator A in $\mathrm{L}^\infty(I)$. Then the following lemma ensures that indeed $\lambda + w_0 = \rho(H+w_0) \geq \rho(H_z + w_0)$.

Lemma 16. *Let J denote a compact interval and let the operators $A = M_A + K_A$ and $B = M_B + K_B$ on $\mathrm{L}^\infty(J)$ be given in terms of multiplication operators M_A, M_B and kernel operators K_A, K_B. Suppose*

$$(M_{A,B}f)(x) = m_{A,B}(x) f(x) \quad \textit{and} \quad (K_{A,B}f)(x) = \int_J k_{A,B}(x,y) f(y)\,\mathrm{d}y \tag{105}$$

with bounded, measurable functions $m_{A,B}$, $k_{A,B}$ satisfying the inequalities $0 \leq m_A \leq m_B$ and $0 \leq k_A \leq k_B$. Then $\rho(A) \leq \rho(B)$.

Proof: Due to the properties of $m_{A,B}$ and $k_{A,B}$, both A and B are positive operators that satisfy $A \leq B$ by assumption, i.e., $Af \leq Bf$ for every $f \geq 0$ (a.e.). An immediate consequence is that $\|A^n\|_\infty \leq \|B^n\|_\infty$ for every $n \in \mathbb{N}$. Thus,

$$\rho(A) = \lim_{n\to\infty} \|A^n\|_\infty^{1/n} \leq \lim_{n\to\infty} \|B^n\|_\infty^{1/n} = \rho(B)\,, \tag{106}$$

compare [Rud91, Thm. 10.13]. (See [Kra72, Thm. 5.3] for a more general result.) □

From now on we need to distinguish the three cases of symmetrizability. For global symmetrizability (90), we consider $\tilde{H}_z$ and want to split it into a part D_z with constant 'diagonals'[8] and a small remainder R_z. Similarly, in the remaining cases, H_z shall be split into a symmetrizable D_z plus a remainder R_z. Both D_z and R_z consist of a multiplication and a kernel part, $D_z = M_{D_z} + K_{D_z}$ and $R_z = M_{R_z} + K_{R_z}$ with

$$\begin{aligned}(M_{D_z}f)(x) &= 1_{J_z}(x)\, m_{D_z}(x)\, f(x)\,,\\ (K_{D_z}f)(x) &= 1_{J_z}(x) \int_{J_z} k_{D_z}(x,y) f(y)\,\mathrm{d}y\,,\end{aligned} \tag{107}$$

and similarly for R_z. The multiplication part is the same in all cases (and constant),

$$m_{D_z} \equiv r(z) - u_1(z)\,, \tag{108}$$

the kernel part differs:

$$k_{D_z}(x,y) = \begin{cases} \tilde{u}\big(z + \frac{1}{2}(x-y), z - \frac{1}{2}(x-y)\big) & \text{for (90)}\\ u\big(z + x - y, z\big) & \text{for (100)}\\ u\big(z + \frac{1}{2}(x-y), z - \frac{1}{2}(x-y)\big) & \text{for (101)}\end{cases} \tag{109}$$

[8]This is a matrix analogy, meaning that, with respect to (107), m_{D_z} is a constant function and $k_{D_z}(x+d, y+d) = k_{D_z}(x,y)$ for all x, $y \in J_z$ and all appropriate d.

In order to determine the appropriate length η of the subintervals, let $\varepsilon > 0$ be given. Due to the continuity of r and u there is an $\eta > 0$ such that, for all $z \in I$ with $J_z \subset I$ and $x,\, y \in J_z$, we have $|(r-u_1)(x) - (r-u_1)(z)| \le \varepsilon/2$ and, depending on the case,

$$\left.\begin{array}{l} |\tilde{u}(x,y) - \tilde{u}(z+\frac{1}{2}(x-y), z-\frac{1}{2}(x-y))| \\ |u(x,y) - u(z+x-y,z)| \\ |u(x,y) - u(z+\frac{1}{2}(x-y), z-\frac{1}{2}(x-y))| \end{array}\right\} \le \frac{\varepsilon}{2|I|}\,. \tag{110}$$

Usually, one will choose η as large as possible. In all cases,

$$|m_{R_z}(x)| \le \frac{\varepsilon}{2} \qquad \text{and} \qquad |k_{R_z}(x,y)| \le \frac{\varepsilon}{2|I|}\,. \tag{111}$$

This leads to

Lemma 17. $\rho(R_z) \le \varepsilon$.

Proof: With (111) one derives

$$\rho(R_z) \le \|R_z\|_\infty \le \|M_{R_z}\|_\infty + \|K_{R_z}\|_\infty\,, \tag{112}$$

where

$$\|K_{R_z}\|_\infty = \sup_{\|f\|_\infty \le 1} \sup_{x\in I} 1_{J_z}(x) \left| \int_{J_z} k_{R_z}(x,y) f(y)\, \mathrm{d}y \right| \le \frac{\varepsilon |J_z|}{2|I|} \le \frac{\varepsilon}{2} \tag{113}$$

and $\|M_{R_z}\|_\infty \le \|m_{R_z}\|_\infty \le \frac{\varepsilon}{2}$. □

We know from Proposition 8 that H is globally symmetrizable if and only if u is of the form (90). An analog for local symmetrizability with the last two cases in (109) is

Proposition 9. *Let D_z be given by* (107), (108), *and either of the last two cases in* (109). *There is a multiplication operator S_z, defined by $S_z f = \exp(s_z) f$, with $s_z \in \mathrm{C}(J_z)$, such that $\tilde{D}_z := S_z D_z S_z^{-1}$ is symmetric if and only if the mutation kernel is of the form* (100) *or* (101), *respectively. Then, the s_z can be chosen as $s_z(x) = -\gamma(z)x + c$ with any $c \in \mathbb{R}$.*

Proof: By Proposition 8, applied to the subinterval J_z, the local symmetry condition is

$$k_{D_z}(x,y) = \exp(\beta_z(x) - \beta_z(y))\, \tilde{k}_{D_z}(x,y)\,. \tag{114}$$

The translational invariance of k_{D_z}, i.e., $k_{D_z}(x+d, y+d) = k_{D_z}(x,y)$, and the symmetry of $\tilde{k}_{D_z}$ imply

$$\beta_z(x) = \gamma(z)x\,, \qquad \tilde{k}_{D_z}(x,y) = h(z, |x-y|)\,, \tag{115}$$

with some $\gamma(z) \in \mathbb{R}$ and some continuous function $h(z,.)\colon [0,\eta] \to \mathbb{R}_{\ge 0}$. For the second case in (109), we have $u(x,y) = k_{D_y}(x,y)$, and, for the last case, $u(x,y) = k_{D_z}(x,y)$ with $z = \frac{1}{2}(x+y)$. Inserting this into (114) and using (115) yields (100), respectively

$$u(x,y) = \exp\Big(\gamma\big(\tfrac{1}{2}(x+y)\big)\tfrac{1}{2}(x-y)\Big) h\big(\tfrac{1}{2}(x+y), |x-y|\big)\,. \tag{116}$$

Since, in the latter case, the last factor is the general form of a symmetric function $\tilde{u}(x,y)$, this is equivalent to (101). In both cases, obviously, $s_z(x) = -\gamma(z)x + c$, with any $c \in \mathbb{R}$, symmetrizes D_z. □

With this we are ready to show

Proposition 10. *If H is globally or locally symmetrizable then*

$$\rho(H_z + w_0) \geq \rho(D_z + w_0) - \varepsilon\,. \tag{117}$$

For the proof, we first recall

Lemma 18. *Let A and B be symmetric operators. Then the triangle inequality holds for the spectral radius, i.e., $\rho(A+B) \leq \rho(A) + \rho(B)$ and $\rho(A-B) \geq |\rho(A) - \rho(B)|$.*

PROOF: For symmetric operators A, the spectral radius $\rho(A)$ coincides with the L^2 operator norm

$$\|A\|_2 = \sup_{0 \neq f \in \mathrm{L}^2(J)} \frac{\|Af\|_2}{\|f\|_2}\,, \tag{118}$$

see, e.g., [Heu92, Thms. 29.5, 112.6], from which the claim follows. □

PROOF OF PROPOSITION 10: For operators K with kernel k, define the symmetric kernel $k^\wedge(x,y) = k(x,y) \wedge k(y,x)$, where $x \wedge y$ denotes the minimum of x and y, and let $K^\wedge$ denote the corresponding kernel operator. If A is the sum of a multiplication operator M_A and a kernel operator K_A, let $A^\wedge = M_A + (K_A)^\wedge$. For the sake of brevity, we now treat both global and local symmetrizability simultaneously. To this end, let $\tilde{D}_z = D_z$, $\tilde{R}_z = R_z$ for global and $\tilde{D}_z$, $\tilde{R}_z$ as in Proposition 9 for local symmetrizability. As an exception, $\tilde{H}_z$ is defined as $(\tilde{H})_z = P_z S H S^{-1} P_z$ for global symmetrizability with S from Proposition 8. Then, since similarity transforms do not change the spectrum, Lemma 16 leads to

$$\rho\big(H_z + w_0\big) = \rho\big(\tilde{H}_z + w_0\big) = \rho\big(\tilde{D}_z + w_0 + \tilde{R}_z\big) \geq \rho\big(\tilde{D}_z + w_0 + (\tilde{R}_z)^\wedge\big)\,. \tag{119}$$

Both $\tilde{D}_z$ and $(\tilde{R}_z)^\wedge$ are symmetric, so Lemma 18 applies and

$$\rho\big(\tilde{D}_z + w_0 + (\tilde{R}_z)^\wedge\big) \geq \rho\big(\tilde{D}_z + w_0\big) - \rho\big((\tilde{R}_z)^\wedge\big) \geq \rho\big(\tilde{D}_z + w_0\big) - \varepsilon = \rho\big(D_z + w_0\big) - \varepsilon\,, \tag{120}$$

which, together with (119), proves the claim. □

With the following proposition we have a lower bound for $\rho(D_z + w_0)$ and are finally able to prove Theorem 8.

Proposition 11. *Let K be the symmetric kernel operator given by*

$$(Kf)(x) = \int_J h(|x-y|) f(y)\,\mathrm{d}y\,, \tag{121}$$

where J is a compact interval of length η and $0 \leq h \in \mathrm{C}([0,\eta])$. Then

$$\rho(K) \geq 2\int_0^{\frac{\eta}{2}} h(\xi)\,\mathrm{d}\xi - 2\int_0^{\frac{\eta}{2}} \int_\xi^{\frac{\eta}{2}} |h(\zeta) - h(\eta-\zeta)|\,\mathrm{d}\zeta\,\mathrm{d}\xi\,. \tag{122}$$

PROOF: Let $\hat{h}(\xi) = h(\xi \wedge (\eta - \xi))$ for $\xi \in [0, \eta]$ and

$$(\hat{K}f)(x) = \int_J \hat{h}(|x-y|) f(y)\, \mathrm{d}y\,. \tag{123}$$

It is easy to see that $\rho(\hat{K}) = 2\int_0^{\frac{\eta}{2}} h(\xi)\,\mathrm{d}\xi$ and that this is an eigenvalue with any constant function as eigenfunction.[9] Since K and $\hat{K}$ are symmetric, Lemma 18 yields

$$\rho(K) \geq \rho(\hat{K}) - \rho(\hat{K} - K)\,. \tag{124}$$

The operator $\hat{K} - K$ can be regarded as a Hille–Tamarkin operator from $\mathrm{L}^q(J)$ to $\mathrm{L}^p(J)$, with any $1 \leq p,\ q \leq \infty$. So, its spectral radius is bounded by the appropriate Hille–Tamarkin norm $|\!|.|\!|_{pq}$. For $p = 1$ and $q = \infty$ this leads to

$$\begin{aligned}\rho(\hat{K} - K) \leq |\!|\hat{K} - K|\!|_{1\infty} &= \int_J \int_J \big|\hat{h}(|x-y|) - h(|x-y|)\big|\,\mathrm{d}y\,\mathrm{d}x \\ &= 2\int_0^{\frac{\eta}{2}} \int_\xi^{\frac{\eta}{2}} \big|h(\zeta) - h(\eta-\zeta)\big|\,\mathrm{d}\zeta\,\mathrm{d}\xi\,,\end{aligned} \tag{125}$$

from which the claim follows. □

PROOF OF THEOREM 8: Let $\varepsilon > 0$ be given and $\eta > 0$ be chosen as above. Then we have, in any of the three cases and for every compact interval $J_z \subset I$ as above,

$$\begin{aligned}\lambda + w_0 = \rho(H + w_0) &\geq \rho(H_z + w_0) \geq \rho(D_z + w_0) - \varepsilon \\ &\geq r(z) - u_1(z) + 2\int_0^{\frac{\eta}{2}} h(z,\xi)\,\mathrm{d}\xi - 2\int_0^{\frac{\eta}{2}} \int_\xi^{\frac{\eta}{2}} |h(z,\zeta) - h(z,\eta-\zeta)|\,\mathrm{d}\zeta\,\mathrm{d}\xi - \varepsilon + w_0 \\ &= r(z) - g_\varepsilon(z) + w_0\end{aligned}$$

due to Lemma 16 and Propositions 10 and 11. □

3.3 An exact limit

In Sections 3.1 and 3.2, upper and lower bounds for the mean fitness have been derived. Here, we discuss the question whether there is a limit in which these bounds converge towards each other and establish a simple maximum principle—analogously to the mutation class limit of the model with discrete genotypes. More precisely, we introduce a parameter $\nu \geq 1$, which we let go to infinity, and replace the mutation kernel u by u_ν, where $\nu = 1$ reproduces u. In the most general case, we want this limit to have the following properties.

[9] To some extent, K and $\hat{K}$ are operator analogs of a Toeplitz and a circulant matrix, see, e.g., [Gra01].

(E1) The u_ν remain globally symmetrizable, such that in each case the upper bound from Theorem 7 is valid.[10]

(E2) Theorem 8 is applicable for every ν and the supremum of the lower bounds converges to the limit of the upper bounds.

It is not clear whether there is a solution to this general case, which I will therefore leave for future work. However, things get much simpler if one adds the following additional assumption, which is also biologically desirable.

(E3) The total mutation rates $\int_I u_\nu(y,x)\,\mathrm{d}y$ remain constant, i.e., equal to $u_1(x)$.

This requires $u_\nu(x,y)$ to be proportional to $u(A_{\nu,y}(x),y)$ with an affine transformation $A_{\nu,y}\colon x \mapsto \alpha(\nu,y)x + \delta(\nu,y)$ for which $A_{\nu,y}^{-1}(I) \subset I$, so in particular $\alpha(\nu,y) \geq 1$. Since we are only interested in the limit $\nu \to \infty$, the scale factor α can be chosen proportional to ν, i.e., $\alpha(\nu,y) = \nu\,\alpha(y)$. Further, we want $A_{\nu,y}(y) = y$, thus $\delta(\nu,y) = -(\nu\,\alpha(y) - 1)y$. So our candidate is

$$u_\nu(x,y) = \nu\, u\big(y + \nu\,\alpha(y)(x-y), y\big)\,, \tag{126}$$

where we set $u(x,y) = 0$ for $(x,y) \notin I \times I$.

Next, due to (E1), all u_ν must be of the form (90). This implies a constant $\alpha(y)$, without loss of generality $\alpha \equiv 1$, linearity of β, $\beta(x) = \gamma\, x$, and translational invariance of $\tilde{u}$, $\tilde{u}(x,y) = h(|x-y|)$. Thus

$$u(x,y) = \exp(\gamma\,(x-y))\,h(|x-y|)\,, \tag{127}$$

$$u_\nu(x,y) = \nu \exp(\nu\,\gamma\,(x-y))\,h(\nu\,|x-y|)\,. \tag{128}$$

This is, of course, a strong restriction since the shape h and the exponential asymmetry factor γ do not depend on the source genotype y. However, (127) describes a subclass of the random-walk mutation model introduced by Crow and Kimura [Cro64], which received considerable attention, see also [Bür00, IV.2].

Given an $\varepsilon > 0$, the largest possible value of η would, in general, depend on ν as well, cf. (110). Due to the translational invariance of u in this case, however, it is solely determined by r and u_1. Thus, we can choose ε_ν, with $\varepsilon_\nu \to 0$ as $\nu \to \infty$, in a way that the corresponding largest value of η, denoted by η_ν, is not smaller than $|I|/\nu$. Then, for every z with $I_{z,\nu} = [z - |I|/\nu, z + |I|/\nu] \subset I$ and every $\eta \in [|I|/\nu, \eta_\nu]$,

$$2\int_0^{\frac{\eta}{2}} \nu\, h(\nu\xi)\,\mathrm{d}\xi \equiv \int_I \sqrt{u(x,z)u(z,x)}\,\mathrm{d}x \tag{129}$$

and

$$\begin{aligned}\int_0^{\frac{\eta}{2}}\int_\xi^{\frac{\eta}{2}} \nu\,|h(\nu\zeta) - h(\nu\eta - \nu\zeta)|\,\mathrm{d}\zeta\,\mathrm{d}\xi &= \int_0^{\frac{\eta}{2}}\int_\xi^{\frac{\eta}{2}} \nu\, h(\nu\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\xi = \int_0^{\frac{\eta}{2}}\int_{\nu\xi}^{\nu\frac{\eta}{2}} h(\zeta)\,\mathrm{d}\zeta\,\mathrm{d}\xi \\ &= \int_0^{\nu\frac{\eta}{2}}\int_0^{\frac{\zeta}{\nu}} h(\zeta)\,\mathrm{d}\xi\,\mathrm{d}\zeta = \int_0^{\nu\frac{\eta}{2}} \frac{\zeta}{\nu} h(\zeta)\,\mathrm{d}\zeta \leq \frac{|I|}{\nu}\int_0^{|I|} h(\zeta)\,\mathrm{d}\zeta \to 0\,.\end{aligned} \tag{130}$$

[10] Alternatively, one may try to find an upper bound based on weaker conditions.

Accordingly,

$$\lim_{\nu\to\infty} \lambda_\nu \geq \lim_{\nu\to\infty} \Big(\sup_{I_{z,\nu}\subset I} \big(r(z) - g(z)\big) + \varepsilon_\nu \Big) = \sup_{z\in I} \big(r(z) - g(z)\big) \tag{131}$$

with g being the limit $\nu \to \infty$ from (98), so (E2) is fulfilled. This, together with (E1), establishes—in this restricted setting—a simple maximum principle for the mean fitness,

$$\lim_{\nu\to\infty} \lambda_\nu = \sup_{z\in I} \big(r(z) - g(z)\big) . \tag{132}$$

Note that the function g, however, is constant due to (E3), (129), and (130).

In any case, if (132) holds and the equilibrium can be characterized by (96) in terms of the ancestral distribution, g again has the interpretation of a mutational loss function. This can be seen by following the arguments made for the model with discrete genotypes in Proposition I.1. Starting from (96), one first shows that $\hat{r}_\nu = \int_I r(x) a_\nu(x)\,\mathrm{d}x \to r(\hat{x})$, if the maximum in (132) is uniquely attained at $\hat{x}$. Then, one considers the mutational loss

$$g_\nu = \hat{r}_\nu - \bar{r}_\nu = \int_I \int_I \sqrt{a_\nu(x)} \sqrt{u_\nu(x,y)\, u_\nu(y,x)} \sqrt{a(y)}\, \mathrm{d}x\, \mathrm{d}y \tag{133}$$

with $\bar{r}_\nu = \lambda_\nu$, cf. Section I.2.5. Since $\bar{r}_\nu \to r(\hat{x}) - g(\hat{x})$ and $\hat{r}_\nu \to r(\hat{x})$, it follows that $g_\nu \to g(\hat{x})$.

3.4 Numerical tests

In the previous section, a limiting procedure was defined in which the upper and lower bounds derived in Sections 3.1 and 3.2 could be shown to converge towards each other and establish a simple maximum principle for the equilibrium mean fitness (132). This, however, was restricted to a special case (127), which is quite unsatisfactory. It is therefore interesting to check numerically if there is hope to generalize these results at least to the cases with a locally symmetrizable mutation kernel discussed in Section 3.2. With the results from Section 2, such a numerical treatment rests on solid grounds, since the continuous models in question can be approximated by discrete ones arbitrarily well and one does not have to fear numerical artefacts.

Let us consider a standard example first, the COA model with quadratic fitness,

$$r(x) = -x^2 , \tag{134}$$

and a Gaussian mutant distribution, i.e., u as defined in (127) with

$$h(|x-y|) = \mu \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|x-y|^2}{2\sigma^2} \right) . \tag{135}$$

Then the multiplication by $\exp(\gamma(x-y))$ in (127) simply leads to a shift in the Gaussian and a modification of its normalization. In this case, the function g is constant (as noted above) and evaluates to

$$g(x) \equiv \exp(\tfrac{1}{2}\gamma^2\sigma^2) - 1 . \tag{136}$$

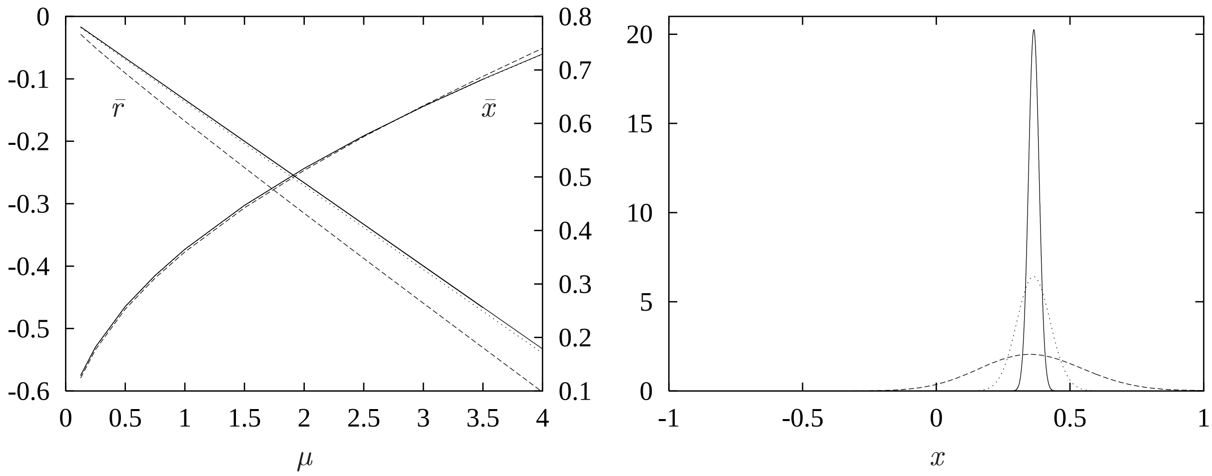

Figure 1: The COA model on $I = [-1, 1]$ with quadratic fitness (134) and a Gaussian mutant distribution (135) with $\sigma = 0.05$ and $\gamma = 10$. The left panel compares the equilibrium mean fitness $\bar{r} = \lambda$ (left axis) and mean genotype $\bar{x} = \int_I x\, p(x)\,\mathrm{d}x$ (right axis) of different values of ν to the predictions according to (132), the maximum principle (solid line). The right panel shows the genotype distribution at $\mu = 1$ getting narrower for increasing values of ν. Dashed lines refer to $\nu = 1$, dotted lines to $\nu = 10$, and solid lines to $\nu = 100$. The latter are, in the left panel, indistinguishable from the predictions of (132).

Some results for $I = [-1, 1]$, $\sigma = 0.05$, and $\gamma = 10$ are shown in Figure 1. The maximum principle predicts the mean fitness $\bar{r} = \lambda$ and the mean genotype $\bar{x} = \int_I x\, p(x)\,\mathrm{d}x$ already with good accuracy for $\nu = 10$, and almost perfectly for $\nu = 100$. For $\nu = 1$ and $\bar{r}$, the agreement is less satisfactory but still qualitatively right.

As a second example, let us again consider quadratic fitness (134), but a locally symmetrizable mutation kernel of the form (100) with

$$\gamma(y) = -10(y - \tfrac{1}{2}) \qquad \text{and} \qquad h(y, |x-y|) = \mu \frac{|x-y|}{2\sigma^2} \exp\left(-\frac{|x-y|^2}{2\sigma^2}\right). \tag{137}$$

Here, the function g involves the error function,

$$g(x) = \sqrt{\frac{\pi}{2}}\, |\gamma(x)|\, \sigma \exp\left(\frac{\gamma(x)^2\sigma^2}{2}\right) \operatorname{erf}\left(\frac{|\gamma(x)|\,\sigma}{\sqrt{2}}\right) - 1\,. \tag{138}$$

Figure 2 shows some results for $I = [-1, 1]$ and $\sigma = 0.05$. In this case, the predictions have the same accuracy as in the previous example, besides for $\nu = 1$ and $\bar{r}$, where they are yet less accurate.

As a last example, Figure 3 shows results for the COA model with the mutation kernel from the previous example (137), hence also the g from (138), but the fitness function

$$r(x) = \frac{1 - \tanh(20(x^2 - 0.1))}{1 + \tanh(2)}\,, \tag{139}$$

depicted in the right panel, again for $I = [-1, 1]$ and $\sigma = 0.05$. The steep cline of the fitness function leads to reduced accuracy for both $\bar{r}$ and $\bar{x}$. But, again, for $\nu = 100$ a difference to the predictions can hardly be seen.

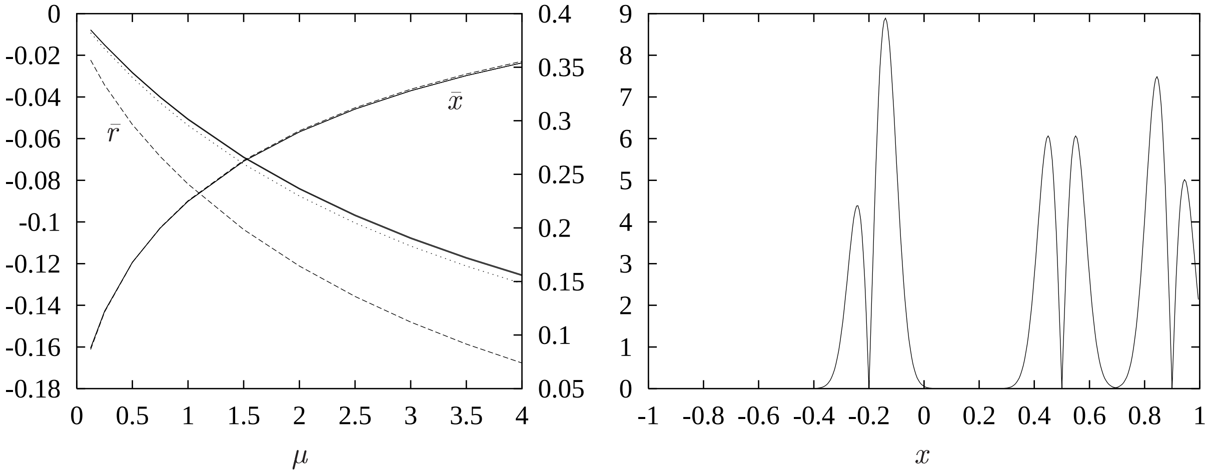

Figure 2: The COA model on $I = [-1, 1]$ with quadratic fitness (134) and a mutation kernel of the form (100) with (137) and $\sigma = 0.05$. The left panel compares the equilibrium mean fitness $\bar{r}$ (left axis) and mean genotype $\bar{x}$ (right axis) of different values of ν to the predictions of the maximum principle (solid line). Dashed lines refer to $\nu = 1$, dotted lines to $\nu = 10$. Again, the case $\nu = 100$ is indistinguishable from the predictions of (132). The right panel shows the mutant distributions $u(x, y)$ for $\nu = 1$ and $y = -0.2$, 0.5, and 0.9, for which the effect of the different asymmetry parameters $\gamma(-0.2) = 7$, $\gamma(0.5) = 0$, and $\gamma(0.9) = -4$ is clearly visible.

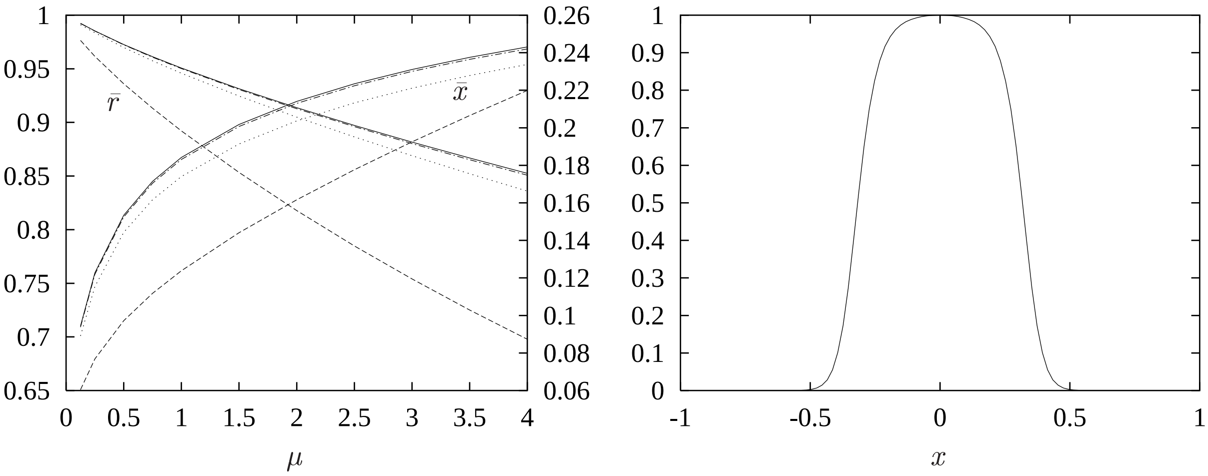

Figure 3: The COA model on $I = [-1, 1]$ with the fitness function (139), shown in the right panel, and a mutation kernel of the form (100) with (137) and $\sigma = 0.05$. The left panel compares the equilibrium mean fitness $\bar{r}$ (left axis) and mean genotype $\bar{x}$ (right axis) of different values of ν to the predictions of the maximum principle (solid line). Dashed lines refer to $\nu = 1$, dotted lines to $\nu = 10$, and dashed-dotted lines to $\nu = 100$.

All these examples make it tempting to conjecture that also for local symmetrizability—at least with (100)—the maximum principle becomes exact in the limit $\nu \to \infty$.

III

Models for unequal crossover

1 The unequal crossover model

This chapter is mainly concerned with a class of models for unequal crossover (UC), originally investigated by Shpak and Atteson [Shp02] for discrete time, which is built on preceding work by Ohta [Oht83] and Walsh [Wal87] (see [Shp02] for further references). Starting from their results, we prove various existence and uniqueness theorems and analyze the convergence properties, both in discrete and in continuous time. In this model class, one considers individuals whose genetic sequences contain a section with repeated units. These may vary in number $i \in \mathbb{N}_0$, where $i = 0$ is explicitly allowed, corresponding to no unit being present (yet). The composition of these sections (with respect to mutations that might have occurred) and the rest of the sequence are ignored.

In the course of time, recombination events take place in which a random pair of individuals is formed and their respective sections are randomly aligned, possibly imperfectly with 'overhangs'. Then, both sequences are cut at a common position between two building blocks and their right (or left) fragments are interchanged. This so-called unequal crossover is schematically depicted in Figure 1. Obviously, the total number of relevant units is conserved in each event.

We assume the population size to be (effectively) infinite.[1] Then, the population is described by a probability measure $\boldsymbol{p} \in \mathcal{M}_1^+(\mathbb{N}_0)$, which we identify with an element $\boldsymbol{p} = (p_k)_{k \in \mathbb{N}_0}$ in the appropriate subset of $\ell^1(\mathbb{N}_0)$. Since we will not consider any genotype space other than $\mathbb{N}_0$ in this chapter, reference to it will be omitted in the sequel. These spaces are complete in the metric derived from the usual ℓ^1 norm, which is the same as the total variation norm here. The metric is denoted by

$$d(\boldsymbol{p}, \boldsymbol{q}) = \|\boldsymbol{p} - \boldsymbol{q}\|_1 = \sum_{k \geq 0} |p_k - q_k| \,. \qquad (1)$$

With this notation, the above process is described by the recombinator

$$\mathcal{R}(\boldsymbol{p})_i = \frac{1}{\|\boldsymbol{p}\|_1} \sum_{j,k,\ell \geq 0} T_{ij,k\ell}\, p_k\, p_\ell \,. \qquad (2)$$

Here, $T_{ij,k\ell} \geq 0$ denotes the probability that a pair (k, ℓ) turns into (i, j), so, for normalization, we require

$$\sum_{i,j \geq 0} T_{ij,k\ell} = 1 \qquad \text{for all } k, \ell \in \mathbb{N}_0 \,. \qquad (3)$$

[1] Concerning finite populations, see the remarks in Section 7.

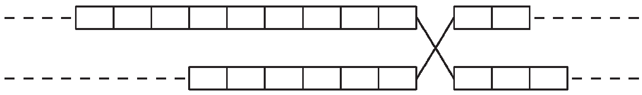

Figure 1: An unequal crossover event as described in the text. Rectangles denote the relevant blocks, while the dashed lines indicate possible extensions with other elements that are disregarded here.

The factor $p_k\, p_\ell$ in (2) describes the probability that a pair (k,ℓ) is formed, i.e., we assume that two individuals are chosen independently from the population. We generally assume that $T_{ij,k\ell}$ is symmetric with respect to both index pairs, $T_{ij,k\ell} = T_{ji,k\ell} = T_{ij,\ell k}$, which is reasonable. Then, the summation over j represents the breaking-up of the pairs after the recombination event. These two ingredients of the dynamics constitute what is known as (*instant*) *mixing* and are responsible for the quadratic nature of the iteration process.

As mentioned above, we will only consider processes that conserve the total copy number in each event, i.e., $T^{(q)}_{ij,k\ell} > 0$ for $i+j = k+\ell$ only. Together with the normalization (3) and the symmetry condition from above, this yields the weaker condition

$$\sum_{i\geq 0} i\, T_{ij,k\ell} = \sum_{i,j\geq 0} \frac{i+j}{2}\, T_{ij,k\ell} = \frac{k+\ell}{2}\,, \tag{4}$$

which implies conservation of the mean copy number in the population,

$$\sum_{i\geq 0} i\, \mathcal{R}(\boldsymbol{p})_i = \sum_{i,j,k,\ell\geq 0} i\, T^{(q)}_{ij,k\ell}\, p_k\, p_\ell = \sum_{k,\ell\geq 0} \frac{k+\ell}{2}\, p_k\, p_\ell = \sum_{k\geq 0} k\, p_k\,. \tag{5}$$

Condition (3) and the presence of the prefactor $1/\|\boldsymbol{p}\|_1$ in (2) make $\mathcal{R}$ norm non-increasing and positive homogeneous of degree 1, i.e., $\mathcal{R}$ satisfies $\|\mathcal{R}(\boldsymbol{x})\|_1 \leq \|\boldsymbol{x}\|_1$ and $\mathcal{R}(a\boldsymbol{x}) = |a|\mathcal{R}(\boldsymbol{x})$, for $\boldsymbol{x} \in \ell^1$ and $a \in \mathbb{R}$. Furthermore, $\mathcal{R}$ is a positive operator with $\|\mathcal{R}(\boldsymbol{x})\|_1 = \|\boldsymbol{x}\|_1$ for all positive elements $\boldsymbol{x} \in \ell^1$. Thus, it is guaranteed that $\mathcal{R}$ maps $\mathcal{M}_r^+$, the space of positive measures of mass r, which is complete in the topology induced by the norm $\|.\|_1$, i.e., by the metric d from (1), into itself. (For $r = 1$, of course, the prefactor is redundant but ensures numerical stability of an iteration with $\mathcal{R}$.)

Given an initial configuration $\boldsymbol{p}_0 = \boldsymbol{p}(t=0)$, the dynamics may be taken in discrete time steps, with subsequent generations,

$$\boldsymbol{p}(t+1) = \mathcal{R}(\boldsymbol{p}(t))\,, \qquad t \in \mathbb{N}_0\,. \tag{6}$$

Our treatment of this case will be set up in a way that also allows for a generalization of the results to the analogous process in continuous time, where generations are overlapping,

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{p}(t) = \varrho\,(\mathcal{R} - \mathbb{1})(\boldsymbol{p}(t))\,, \qquad t \in \mathbb{R}_{\geq 0}\,. \tag{7}$$

Obviously, the (positive) parameter ϱ in (7) only leads to a rescaling of the time t. We therefore choose $\varrho = 1$ without loss of generality. Furthermore, it is easily verified that the fixed points of (6) are in one-to-one correspondence with the equilibria[2] of (7).

[2]In the sequel, we will use the term fixed point for both discrete and continuous dynamics.

In the UC model, one distinguishes 'perfect' alignments, in which each unit in the shorter sequence has a partner in the longer sequence, and 'imperfect' alignments, with 'overhangs' of the shorter sequence. The first are taken to be equally probable, the latter penalized by a factor q^d relative to the first, where $q \in [0,1]$ is a model parameter and d the length of the overhang (at most the entire length of the shorter sequence; in the example of Figure 1, we have $d = 1$). In the extreme case $q = 0$, only perfect alignments may occur, whereas for $q = 1$ overhangs are not penalized at all. For obvious reasons, the first case is dubbed *internal UC*, the second *random UC* [Shp02].

In compact notation, this leads to the transition probabilities

$$T^{(q)}_{ij,k\ell} = C^{(q)}_{k\ell}\, \delta_{i+j,k+\ell}\, (1 + \min\{k,\ell,i,j\})\, q^{0 \vee (k\wedge\ell - i\wedge j)}\,, \tag{8}$$

where $k \vee \ell := \max\{k,\ell\}$, $k \wedge \ell := \min\{k,\ell\}$, and $0^0 = 1$. The $C^{(q)}_{k\ell}$ are chosen such that (3) holds, i.e., $\sum_{i,j\geq 0} T^{(q)}_{ij,k\ell} = 1$, and are hence symmetric in k and ℓ. Explicitly, they read (see also [Shp02, Sec. 2.1])

$$C^{(q)}_{k\ell} = \frac{(1-q)^2}{(k\wedge\ell+1)(|k-\ell|+1)(1-q)^2 + 2q(k\wedge\ell - (k\wedge\ell+1)q + q^{k\wedge\ell+1})}\,. \tag{9}$$

Note further that the total number of units is indeed conserved in each event and that the process is symmetric within both pairs. Hence (4) is satisfied.

The aim of this chapter is to find answers to the following questions:

1. Are there fixed points of the dynamics?
2. Given the mean copy number m, is there a unique fixed point?
3. If so, under which conditions and in which sense does an initial distribution converge to this fixed point?

Of course, the trivial fixed point with $p_0 = 1$ and $p_k = 0$ for $k > 0$ always exists, which we generally exclude from our considerations. But even then, the answer to the first question is positive for general operators of the form (2) satisfying (3) and some rather natural further condition. This is discussed in Section 2. For the extreme cases $q = 0$ (internal UC) and $q = 1$ (random UC), fixed points are known explicitly for every m and it has been conjectured that, under mild conditions, also questions 2 and 3 can be answered positively for all values of $q \in [0,1]$ [Shp02]. Indeed, for both extreme cases, norm convergence of the population distribution to the fixed points can be shown, which is done in Sections 3 and 5, respectively. Since the dynamical systems involved are infinite dimensional, a careful analysis of compactness properties is needed for rigorous answers. The proofs for $q = 1$ are based on alternative representations of probability measures via generating functions, presented in Section 4. For the intermediate parameter regime, we can only show that there exists a fixed point for every m, but neither its uniqueness nor convergence to it, see Section 6. Some remarks in Section 7 conclude this chapter.

2 Existence of fixed points

Let us begin by stating the following general fact.

Proposition 1. *If the recombinator $\mathcal{R}$ of* (2) *satisfies* (3)*, then the global Lipschitz condition*

$$\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 \leq C\|\boldsymbol{x} - \boldsymbol{y}\|_1 \tag{10}$$

is fulfilled, with constant $C = 3$ on ℓ^1, respectively $C = 2$ if $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}_r$.

Proof: Let $\boldsymbol{x}, \boldsymbol{y} \in \ell^1$ be non-zero (otherwise the statement is trivial). Then

$$\begin{aligned}
\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 &= \sum_{i\geq 0} \left| \sum_{j,k,\ell \geq 0} T_{ij,k\ell} \left(\frac{x_k\, x_\ell}{\|\boldsymbol{x}\|_1} - \frac{y_k\, y_\ell}{\|\boldsymbol{y}\|_1} \right) \right| \\
&\leq \sum_{k,\ell\geq 0} \left| \frac{x_k\, x_\ell}{\|\boldsymbol{x}\|_1} - \frac{y_k\, y_\ell}{\|\boldsymbol{y}\|_1} \right| \sum_{i,j\geq 0} T_{ij,k\ell} = \sum_{k,\ell\geq 0} \left| \frac{x_k\, x_\ell}{\|\boldsymbol{x}\|_1} - \frac{x_k\, y_\ell}{\|\boldsymbol{x}\|_1} + \frac{x_k\, y_\ell}{\|\boldsymbol{x}\|_1} - \frac{y_k\, y_\ell}{\|\boldsymbol{y}\|_1} \right| \\
\leq \sum_{k,\ell\geq 0} \left(\frac{|x_k|}{\|\boldsymbol{x}\|_1} |x_\ell - y_\ell| + |y_\ell| \left| \frac{x_k}{\|\boldsymbol{x}\|_1} - \frac{y_k}{\|\boldsymbol{y}\|_1} \right| \right) &= \|\boldsymbol{x} - \boldsymbol{y}\|_1 + \frac{1}{\|\boldsymbol{x}\|_1} \big\| \|\boldsymbol{y}\|_1 \boldsymbol{x} - \|\boldsymbol{x}\|_1 \boldsymbol{y} \big\|_1 .
\end{aligned}$$

The last term becomes

$$\frac{1}{\|\boldsymbol{x}\|_1} \big\| \|\boldsymbol{y}\|_1 \boldsymbol{x} - \|\boldsymbol{x}\|_1 \boldsymbol{y} \big\|_1 = \frac{1}{\|\boldsymbol{x}\|_1} \big\| (\|\boldsymbol{y}\|_1 - \|\boldsymbol{x}\|_1)\boldsymbol{x} + \|\boldsymbol{x}\|_1 (\boldsymbol{x} - \boldsymbol{y}) \big\|_1 \leq 2\|\boldsymbol{x} - \boldsymbol{y}\|_1 , \tag{11}$$

from which $\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 \leq 3\|\boldsymbol{x} - \boldsymbol{y}\|_1$ follows for $\boldsymbol{x}, \boldsymbol{y} \in \ell^1$. If $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}_r$, the above calculation simplifies to $\|\mathcal{R}(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{y})\|_1 \leq 2\|\boldsymbol{x} - \boldsymbol{y}\|_1$. □

In continuous time, this is a sufficient condition for the existence and uniqueness of a solution of the initial value problem (7), cf. [Ama90, Thms. 7.6 and 10.3]. Another useful notion in this respect is the following.

Definition 1 [Ama90, Sec. 18]. *Let Y be an open subset of a Banach space E and let $f\colon Y \to E$ satisfy a* (*local*) *Lipschitz condition. A continuous function L from $X \subset Y$ to $\mathbb{R}$ is called a* Lyapunov function *for the initial value problem*

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}(t) = f(\boldsymbol{x}(t)) , \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0 \in X , \tag{12}$$

if the orbital derivative $\dot{L}(\boldsymbol{x}_0) := \liminf_{t\to 0^+} \frac{1}{t}\big(L(\boldsymbol{x}(t)) - L(\boldsymbol{x}_0)\big)$ *satisfies*

$$\dot{L}(\boldsymbol{x}_0) \leq 0 \tag{13}$$

for all initial conditions $\boldsymbol{x}_0 \in X$.

If further $\dot{L}(\boldsymbol{x}_{\mathrm{F}}) = 0$ for a single fixed point $\boldsymbol{x}_{\mathrm{F}}$ only, then L is called a *strict* Lyapunov function. If a Lyapunov function exists, we have

Theorem 1 [Ama90, Thm. 17.2 and Cor. 18.4]. *With the notation of Definition* 1, *assume that there is a Lyapunov function L, that the set X is closed, and that, for an initial condition $\boldsymbol{x}_0 \in X$, the set $\{\boldsymbol{x}(t) : t \in \mathbb{R}_{\geq 0}, \boldsymbol{x}(t)$ exists$\}$ is relatively compact in X. Then, $\boldsymbol{x}(t)$ exists for all $t \geq 0$ and*

$$\lim_{t\to\infty} \operatorname{dist}(\boldsymbol{x}(t), X_L) = 0\,, \tag{14}$$

where X_L denotes the largest invariant subset of $\{\boldsymbol{x} \in X : \dot{L}(\boldsymbol{x}) = 0\}$ (in forward and backward time) and $\operatorname{dist}(\boldsymbol{x}, X_L) = \inf_{\boldsymbol{y}\in X_L} \|\boldsymbol{x} - \boldsymbol{y}\|$.

Obviously, if L is a strict Lyapunov function, we have $X_L = \{\boldsymbol{x}_{\mathrm{F}}\}$ and this theorem implies $d(\boldsymbol{x}(t), \boldsymbol{x}_{\mathrm{F}}) \to 0$ as $t \to \infty$.

Returning to the original question of the existence of fixed points, we now recall the following facts, compare [Bil99, Shi96] for details.

Proposition 2 [Yos80, Cor. to Thm. V.1.5]. *Assume the sequence $\big(\boldsymbol{p}^{(n)}\big) \subset \mathcal{M}_1^+$ to converge in the weak-$*$ topology (i.e., pointwise, or vaguely) to some $\boldsymbol{p} \in \mathcal{M}_1^+$, i.e.,*

$$\lim_{n\to\infty} p_k^{(n)} = p_k \quad \textit{for all } k \in \mathbb{N}_0\,, \qquad \textit{with } p_k \geq 0 \textit{ and } \textstyle\sum_{k\geq 0} p_k = 1\,. \tag{15}$$

Then it also converges weakly (in the probabilistic sense) and in total variation, i.e., $\lim_{n\to\infty} \|\boldsymbol{p}^{(n)} - \boldsymbol{p}\|_1 = 0$.

Proposition 3. *Assume that the recombinator $\mathcal{R}$ from* (2) *satisfies* (3) *and has a convex, weak-$*$ closed invariant set $M \subset \mathcal{M}_1^+$, i.e., $\mathcal{R}(M) \subset M$, that is tight, i.e., for every $\varepsilon > 0$ there is an $m \in \mathbb{N}_0$ such that $\sum_{k\geq m} p_k < \varepsilon$ for every $\boldsymbol{p} \in M$. Then $\mathcal{R}$ has a fixed point in M.*

Proof: Prohorov's theorem [Shi96, Thm. III.2.1] states that tightness and weak-$*$ relative compactness are equivalent (see also [Bil99, Chs. 1.1 and 1.5]). In our case, M is tight and weak-$*$ closed, therefore, due to Proposition 2, norm compact. Further, M is convex and $\mathcal{R}$ is (norm) continuous by Proposition 1. Thus, the claim follows from the Leray–Schauder–Tychonov fixed point theorem [Ree80, Thm. V.19]. □

With respect to the UC model, we will indeed find such compact invariant subsets.

3 Internal unequal crossover

After these preliminaries, let us begin with the case of internal UC with perfect alignment only, i.e., $q = 0$ in (8). This case is the simplest because, in each recombination event, no sequences longer than the participating sequences can be formed. Here, on $\mathcal{M}_1^+$, the recombinator (2) simplifies to

$$\mathcal{R}_0(\boldsymbol{p})_i = \sum_{\substack{k,\ell\geq 0\\ k\wedge\ell\leq i\leq k\vee\ell}} \frac{p_k\, p_\ell}{1 + |k-\ell|}\,. \tag{16}$$

From now on, we write $\mathcal{R}_q$ rather than $\mathcal{R}$ whenever we look at a recombinator with (fixed) parameter q. It is instructive to generalize the notion of reversibility (or detailed balance, compare [Shp02, (4.1)]).

Definition 2. *We call a probability measure $\boldsymbol{p} \in \mathcal{M}_1^+$* reversible *for a recombinator $\mathcal{R}$ of the form* (2) *if, for all $i,\ j,\ k,\ \ell \geq 0$,*

$$T_{ij,k\ell}\, p_k\, p_\ell = T_{k\ell,ij}\, p_i\, p_j\,. \tag{17}$$

The relevance of this concept is evident from the following property.

Lemma 1. *If $\boldsymbol{p} \in \mathcal{M}_1^+$ is reversible for $\mathcal{R}$, it is also a fixed point of $\mathcal{R}$.*

PROOF: Assume $\boldsymbol{p}$ to be reversible. Then, by (3),

$$\mathcal{R}(\boldsymbol{p})_i = \sum_{j,k,\ell\geq 0} T_{ij,k\ell}\, p_k\, p_\ell = \sum_{j,k,\ell\geq 0} T_{k\ell,ij}\, p_i\, p_j = p_i \sum_{j\geq 0} p_j = p_i\,. \tag{18}$$

□

So, in our search for fixed points, we start by looking for solutions of (17). Since, for $q = 0$, forward and backward transition probabilities are simultaneously non-zero only if $\{i,j\} = \{k,\ell\} \subset \{n, n+1\}$ for some n, the components p_k may only be positive on this small set as well. By the following proposition, this indeed characterizes all fixed points $q = 0$.

Proposition 4. *A probability measure $\boldsymbol{p} \in \mathcal{M}_1^+$ is a fixed point of $\mathcal{R}_0$ if and only if its mean copy number $m = \sum_{k\geq 0} k\, p_k$ is finite, $p_{\lfloor m\rfloor} = \lfloor m\rfloor + 1 - m$, $p_{\lceil m\rceil} = m + 1 - \lceil m\rceil$, and $p_k = 0$ for all other k. This includes the case that m is integer and $p_{\lfloor m\rfloor} = p_{\lceil m\rceil} = p_m = 1$.*

PROOF: The 'if' part was stated in [Shp02, Sec. 4.1] and follows easily by insertion into (16) or (17). For the 'only if' part, let i denote the smallest integer such that $p_i > 0$. Then

$$\mathcal{R}(\boldsymbol{p})_i = p_i^2 + 2p_i \sum_{\ell\geq 1} \frac{p_{i+\ell}}{1+\ell} = p_i \left(p_i + p_{i+1} + \sum_{\ell\geq 2} \frac{2}{\ell+1} p_{i+\ell} \right) \leq p_i\,, \tag{19}$$

where the last step follows since $\frac{2}{\ell+1} < 1$ in the last sum, with equality if and only if $p_k = 0$ for all $k \geq i+2$. This implies $m < \infty$ and the uniqueness of $\boldsymbol{p}$ (given m) with the non-zero frequencies as claimed. □

It it possible to analyze the case of internal UC on the basis of the compact sets to be introduced below in Section 4. However, as J. Hofbauer pointed out to us [Hofb], it is more natural to start with a larger compact set to be introduced in (20). Our main result in this section is thus

Theorem 2. *Assume that, for the initial condition $\boldsymbol{p}(0)$ and fixed $r > 1$, the r-th moment exists, $\sum_{k\geq 0} k^r p_k(0) < \infty$. Then $m = \sum_{k\geq 0} k\, p_k(0)$ is finite and, both in discrete and in continuous time, $\lim_{t\to\infty} \|\boldsymbol{p}(t) - \boldsymbol{p}\|_1 = 0$ with the appropriate fixed point $\boldsymbol{p}$ from Proposition* 4.

The proof relies on the following lemma, which slightly modifies and completes the convergence arguments of [Shp02, Sec. 4.1], puts them on rigorous grounds, and extends them to continuous time.

Lemma 2. *Let $r > 1$ be arbitrary, but fixed. Consider the set of probability measures with fixed mean $m < \infty$ and a centered r-th moment bounded by $C < \infty$,*

$$\mathcal{M}^+_{1,m,C} = \{\boldsymbol{p} \in \mathcal{M}^+_1 : \sum_{k\geq 0} k\, p_k = m,\ M_r(\boldsymbol{p}) \leq C\}\,, \tag{20}$$

equipped with (the metric induced by) the total variation norm, where

$$M_s(\boldsymbol{p}) = \sum_{k\geq 0} |k-m|^s\, p_k \tag{21}$$

for $s \in \{1, r\}$. This is a compact and convex space. Both M_1 and M_r satisfy the inequality $M_s(\mathcal{R}_0(\boldsymbol{p})) \leq M_s(\boldsymbol{p})$, with equality if and only if $\boldsymbol{p}$ is a fixed point of $\mathcal{R}_0$. Furthermore, M_1 is a continuous mapping from $\mathcal{M}^+_{1,m,C}$ to $\mathbb{R}_{\geq 0}$ and a Lyapunov function for the dynamics in continuous time.

Proof: Let a sequence $(\boldsymbol{p}^{(n)}) \subset \mathcal{M}^+_{1,m,C}$ be given and consider the random variables $\boldsymbol{f}^{(n)} = (k)_{k\in\mathbb{N}_0}$ on the probability spaces $(\mathbb{N}_0, \boldsymbol{p}^{(n)})$. Their expectation values are equal to m, which, by Markov's inequality [Shi96, p. 599], implies the tightness of the sequence $(\boldsymbol{p}^{(n)})$. Hence, by Prohorov's theorem [Shi96, Thm. III.2.1] (see also [Bil99, Chs. 1.1 and 1.5]), it contains a convergent subsequence $(\boldsymbol{p}^{(n_i)})$ (recall that, by Proposition 2, norm and pointwise convergence are equivalent in this case). Let $\tilde{\boldsymbol{p}} \in \mathcal{M}^+_1$ denote its limit and $\tilde{\boldsymbol{f}} = (k)_{k\in\mathbb{N}_0}$ a random variable on $(\mathbb{N}_0, \tilde{\boldsymbol{p}})$, to which the $\boldsymbol{f}^{(n_i)}$ converge in distribution. Since $r > 1$, the $\boldsymbol{f}^{(n_i)}$ are uniformly integrable by Markov's and Hölder's inequalities. Hence, due to [Kal97, Lem. 3.11], their expectation values, which all equal m, converge to the one of $\tilde{\boldsymbol{f}}$, which is thus m as well. Now consider the random variables $\boldsymbol{g}^{(n_i)} = \tilde{\boldsymbol{g}} = (|k-m|^r)_{k\in\mathbb{N}_0}$ on $(\mathbb{N}_0, \boldsymbol{p}^{(n)})$ and $(\mathbb{N}_0, \tilde{\boldsymbol{p}})$, respectively. The expectation values of the $\boldsymbol{g}^{(n_i)}$ are bounded by C, which, again by [Kal97, Lem. 3.11], is then also an upper bound for the expectation value of $\tilde{\boldsymbol{g}}$ (to which the $\boldsymbol{g}^{(n_i)}$ converge in distribution). This proves the compactness of $\mathcal{M}^+_{1,m,C}$. The convexity is obvious.

With respect to the second statement, consider

$$\begin{aligned} M_s(\mathcal{R}_0(\boldsymbol{p})) &= \sum_{i\geq 0} \sum_{\substack{k,\ell\geq 0\\ k\wedge\ell\leq i\leq k\vee\ell}} \frac{|i-m|^s}{1+|k-\ell|}\, p_k\, p_\ell \\ &= \sum_{k,\ell\geq 0} \frac{p_k\, p_\ell}{1+|k-\ell|}\, \frac{1}{2} \sum_{i=k\wedge\ell}^{k\vee\ell} \left(|i-m|^s + |k+\ell-i-m|^s\right). \end{aligned} \tag{22}$$

For notational convenience, let $j = k+\ell-i$. We now show

$$|i-m|^s + |k+\ell-i-m|^s \leq |k-m|^s + |\ell-m|^s\,. \tag{23}$$

If $\{k,\ell\} = \{i,j\}$, then (23) holds with equality. Otherwise, assume without loss of generality that $k < i \leq j < \ell$. If $m \leq k$ or $m \geq \ell$, we have equality for $s = 1$,[3] but a

[3]This describes the fact that a recombination event between two sequences that are both longer or both shorter than the mean does not change their mean distance to the mean copy number.

strict inequality for $s = r$ due to the convexity of $x \mapsto x^r$. In the remaining cases, the inequality is strict as well. Hence, $M_s(\mathcal{R}_0(\boldsymbol{p})) \leq M_s(\boldsymbol{p})$ with equality if and only if $\boldsymbol{p}$ is a fixed point of $\mathcal{R}_0$, since otherwise the sum in (22) contains at least one term for which (23) holds as a strict inequality.

To see that M_1 is continuous, consider a converging sequence $(\boldsymbol{p}^{(n)})$ in $\mathcal{M}^+_{1,m,C}$ and the random variables $\boldsymbol{h}^{(n)} = (|k-m|)_{k\in\mathbb{N}_0}$ on $(\mathbb{N}_0, \boldsymbol{p}^{(n)})$. As above, the latter are uniformly integrable, from which the continuity of M_1 follows. Since $M_1(\boldsymbol{p})$ is linear in $\boldsymbol{p}$ and thus infinitely differentiable, so is the solution $\boldsymbol{p}(t)$ for every initial condition $\boldsymbol{p}_0 \in \mathcal{M}^+_{1,m,C}$, compare [Ama90, Thm. 9.5 and Rem. 9.6(b)]. Therefore, we have

$$\dot{M}_1(\boldsymbol{p}_0) = \liminf_{t\to 0^+} \frac{M_1(\boldsymbol{p}(t)) - M_1(\boldsymbol{p}_0)}{t} = M_1(\mathcal{R}_0(\boldsymbol{p}_0)) - M_1(\boldsymbol{p}_0) \leq 0\,,$$

again with equality if and only if $\boldsymbol{p}_0$ is a fixed point. Thus, M_1 is a Lyapunov function. □

Proof of Theorem 2. By assumption, the r-th moment of $\boldsymbol{p}(0)$ exists, which is equivalent to the existence of the centered r-th moment by Minkowski's inequality [Shi96, Sec. II.6.6]. This obviously implies the existence of the mean m. By Lemma 2, $\boldsymbol{p}(t) \in P_{\alpha,\delta}$ follows for all $t \geq 0$, directly for discrete time and via a satisfied subtangent condition [Mar76, Thm. VI.2.1] (see also [Ama90, Thm. 16.5]) for continuous time. In the discrete case, due to the compactness of $\mathcal{M}^+_{1,m,C}$, there is a convergent subsequence $(\boldsymbol{p}(t_i))$ with some limit $\boldsymbol{p}$. Now consider the mean distance M_1 to the mean copy number from (21). If $\lim_{t\to\infty} \boldsymbol{p}(t) = \boldsymbol{p}$, we have, due to the continuity of M_1 and $\mathcal{R}_0$,

$$M_1(\mathcal{R}_0(\boldsymbol{p})) = \lim_{t\to\infty} M_1(\mathcal{R}_0(\boldsymbol{p}(t))) = \lim_{t\to\infty} M_1(\boldsymbol{p}(t+1)) = M_1(\boldsymbol{p})\,, \tag{24}$$

thus $\boldsymbol{p}$ is a fixed point by Lemma 2. Otherwise, there are two convergent subsequences $(\boldsymbol{p}(t_i))$, with limit $\boldsymbol{p}$, and $(\boldsymbol{p}(s_i))$, with limit $\boldsymbol{q}$, which satisfy $t_i < s_i < t_{i+1}$. Then, we also have $M_1(\mathcal{R}_0(\boldsymbol{p}(t_i))) \geq M_1(\boldsymbol{p}(s_i))$ and $M_1(\mathcal{R}_0(\boldsymbol{p}(s_i))) \geq M_1(\boldsymbol{p}(t_{i+1}))$, and therefore

$$\begin{aligned} M_1(\boldsymbol{p}) \geq M_1(\mathcal{R}_0(\boldsymbol{p})) &= \lim_{i\to\infty} M_1(\mathcal{R}_0(\boldsymbol{p}(t_i))) \geq \lim_{i\to\infty} M_1(\boldsymbol{p}(s_i)) = M_1(\boldsymbol{q}) \\ \geq M_1(\mathcal{R}_0(\boldsymbol{q})) &= \lim_{i\to\infty} M_1(\mathcal{R}_0(\boldsymbol{p}(s_i))) \geq \lim_{i\to\infty} M_1(\boldsymbol{p}(t_{i+1})) = M_1(\boldsymbol{p})\,. \end{aligned} \tag{25}$$

Thus, both $\boldsymbol{p}$ and $\boldsymbol{q}$ are fixed points by Lemma 2 and hence equal by Proposition 4. In continuous time, the claim follows from Theorem 1 since M_1 is a Lyapunov function by Lemma 2. □

Note that, for $q = 0$, the recombinator can be expressed as $\mathcal{R}_0(\boldsymbol{p})_i = \sum_{j=0}^{i} \pi_{j,i-j}$ in terms of explicit frequencies $\pi_{k,\ell}$ of fragment pairs before concatenation (with copy numbers k and ℓ). However, we have, so far, not been able to use this for a simplification of the above treatment.

4 Alternative probability representations

In the following treatment, we will consider, as an alternative representation for a probability measure $\boldsymbol{p} \in \mathcal{M}_1^+$, the generating function

$$\psi(z) = \sum_{k \geq 0} p_k z^k \,, \tag{26}$$

for which $\psi(1) = \|\boldsymbol{p}\|_1 = 1$ and the radius of convergence is at least 1. We will restrict our discussion to such $\boldsymbol{p}$ for which $\limsup_{k\to\infty} \sqrt[k]{p_k} < 1$. Then, by Hadamard's formula [Rud86, 10.5], the radius of convergence, $\rho(\psi) = 1/\limsup_{k\to\infty} \sqrt[k]{p_k}$, is larger than 1. This is, biologically, no restriction since for any 'realistic' system there are only finitely many non-zero p_k (and thus $\rho(\psi) = \infty$). Mathematically, this condition ensures the existence of all moments and enables us to go back and forth between the probability measure and its generating function, even when looked at $\psi(z)$ only in the vicinity of $z = 1$ (see Proposition 6 below and [Shi96, Sec. II.12]). By abuse of notation, we define the induced recombinator for these generating functions as

$$\mathcal{R}(\psi)(z) = \sum_{k \geq 0} \mathcal{R}(\boldsymbol{p})_k \, z^k \,. \tag{27}$$

In general, with the exception of the case $q = 1$, we do not know any simple expression for $\mathcal{R}(\psi)$ in terms of ψ. Nevertheless, (27) will be central to our further analysis.

It is advantageous to use the local expansion around $z = 1$, written in the form

$$\psi(z) = \sum_{k \geq 0} (k+1) a_k (z-1)^k \,, \tag{28}$$

whose coefficients are given by

$$a_k = \frac{1}{(k+1)!} \psi^{(k)}(1) = \frac{1}{k+1} \sum_{\ell \geq k} \binom{\ell}{k} p_\ell =: \boldsymbol{a}(\boldsymbol{p})_k \geq 0 \,. \tag{29}$$

In particular, $a_0 = 1$ and $a_1 = \frac{1}{2} \sum_{\ell \geq 0} \ell \, p_\ell$. This definition of a_k is size biased, and will become clear from the simplified dynamics for $q = 1$ that results from it. For the sake of compact notation, we use $\boldsymbol{a} = (a_k)_{k \in \mathbb{N}_0}$ both for the coefficients and for the mapping. The coefficients $\boldsymbol{a}$ are elements of the following compact, convex metric space.

Definition 3. *For fixed α and δ with $0 < \alpha \leq \delta < \infty$, let*

$$X_{\alpha,\delta} = \{\boldsymbol{a} = (a_k)_{k \in \mathbb{N}_0} : a_0 = 1,\, a_1 = \alpha,\, 0 \leq a_k \leq \delta^k \text{ for } k \geq 2\} \,. \tag{30}$$

On this space, define the metric

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sum_{k \geq 0} d_k |a_k - b_k| \tag{31}$$

with $d_k = (\gamma/\delta)^k$ for some $0 < \gamma < \frac{1}{3}$.

Obviously, d is indeed a metric and $X_{\alpha,\delta}$ is a convex set, i.e., $\eta\,\boldsymbol{a} + (1-\eta)\boldsymbol{b} \in X_{\alpha,\delta}$ for all $\boldsymbol{a}$, $\boldsymbol{b} \in X_{\alpha,\delta}$ and $\eta \in [0,1]$. Note that we use the same symbol d as in (1) since it will always be clear which metric is meant. The space $X_{\alpha,\delta}$ is naturally embedded in the Banach space (cf. [Wal98, Sec. 24.I])

$$H_{\gamma/\delta} = \{\boldsymbol{x} \in \mathbb{R}^{\mathbb{N}_0} : \|\boldsymbol{x}\| < \infty\} \tag{32}$$

with the norm $\|\boldsymbol{x}\| = \sum_{k\geq 0}(\gamma/\delta)^k |x_k|$, for γ and δ as in Definition 3. In particular, $d(\boldsymbol{a},\boldsymbol{b}) = \|\boldsymbol{a}-\boldsymbol{b}\|$. Further, we have the following two propositions.

Proposition 5. *The space $X_{\alpha,\delta}$ is compact in the metric d of* (31)*.*

Proof: In metric spaces, compactness and sequential compactness are equivalent, compare [Lan93, Thm. II.3.8]. Therefore, let $(\boldsymbol{a}^{(n)})$ be any sequence in $X_{\alpha,\delta}$. By assumption, $a_0^{(n)} \equiv 1$ and $a_1^{(n)} \equiv \alpha$. Furthermore, every element sequence $(a_k^{(n)}) \subset [0,\delta^k]$ has a convergent subsequence. We now inductively define, for every k, a convergent subsequence $(a_k^{(n_{k,i})})$, with limit a_k, such that the indices $\{n_{k,i} : i \in \mathbb{N}\}$ are a subset of the preceding indices $\{n_{k-1,i} : i \in \mathbb{N}\}$. This way, we can proceed to a 'diagonal' sequence $(\boldsymbol{a}^{(n_{i,i})})$. The latter is now shown to converge to $\boldsymbol{a} = (a_k)$, which is obviously an element of $X_{\alpha,\delta}$. To this end, let $\varepsilon > 0$ be given. Choose m large enough such that $\sum_{k>m}(2\gamma)^k < \varepsilon/2$, and then i such that $\sum_{k=0}^m d_k|a_k^{(n_{i,i})} - a_k| < \varepsilon/2$. Then

$$d(\boldsymbol{a}^{(n_{i,i})},\boldsymbol{a}) = \sum_{k=2}^{m} d_k|a_k^{(n_{i,i})} - a_k| + \sum_{k>m} d_k|a_k^{(n_{i,i})} - a_k| < \frac{\varepsilon}{2} + \sum_{k>m}(2\gamma)^k < \varepsilon\,, \tag{33}$$

which proves the claim. □

Proposition 6. *If* $\limsup_{k\to\infty}\sqrt[k]{p_k} < 1$*, the coefficients a_k of* (29) *exist and $\boldsymbol{a}(\boldsymbol{p}) \in X_{\alpha,\delta}$ with $\alpha = \boldsymbol{a}(\boldsymbol{p})_1 = \frac{1}{2}m = \frac{1}{2}\sum_{k\geq 0} k\,p_k$ and some δ. Conversely, if $\boldsymbol{p}(\boldsymbol{a}) \in X_{\alpha,\delta}$ for some α, δ, one has* $\limsup_{k\to\infty}\sqrt[k]{p_k} < 1$.

For a proof, we need the following

Lemma 3. *Let $f_0(z) = \sum_{k\geq 0} c_k z^k$ be a power series with non-negative coefficients c_k and $f_x(z) = \sum_{k\geq 0}\frac{1}{k!}f_0^{(k)}(x)(z-x)^k$ the expansion of f_0 around some $x \in [0,\rho(f_0)[$. Then $\rho(f_0) = x + \rho(f_x)$, including the case that both radii of convergence are infinite.*

Proof: The inequality $\rho(f_x) \geq \rho(f_0) - x$ immediately follows from the theorem of representability by power series [Rud86, Thm. 10.16] since the open disc $B_x(\rho(f_0)-x)$ is entirely included in $B_0(\rho(f_0))$. Consider the power series $f_{xe^{i\varphi}}(z) = \sum_{k\geq 0}\frac{1}{k!}f_0^{(k)}(xe^{i\varphi})(z-xe^{i\varphi})^k$ with arbitrary $\varphi \in [0,2\pi[$. Due to the non-negativity of the c_k, its coefficients fulfill $|f_0^{(k)}(xe^{i\varphi})| \leq \sum_{n\geq k}\frac{n!}{(n-k)!}c_k x^{n-k} = f_0^{(k)}(x)$. With this, Hadamard's formula implies $\rho(f_{xe^{i\varphi}}) \geq \rho(f_x)$. Therefore, f_0 admits an analytic continuation on $B_0(x+\rho(f_x))$, the uniqueness of which follows from the monodromy theorem [Rud86, Thm. 16.16]. The theorem of representability by power series then yields $\rho(f_0) \geq x + \rho(f_x)$, which, together with the opposite inequality above, proves the claim. □

Proof of Proposition 6: The assumption implies $\rho(\psi) > 1$ for ψ from (26). Then, from Lemma 3, we know that $\limsup_{k\to\infty} \sqrt[k]{(k+1)a_k} < \infty$. Since $a_k \leq (k+1)a_k$, also $\limsup_{k\to\infty} \sqrt[k]{a_k} < \infty$, so there is an upper bound δ for $\sqrt[k]{a_k}$ and thus $\boldsymbol{a}(\boldsymbol{p}) \in X_{\alpha,\delta}$. The converse statement follows from (29) and Lemma 3. □

Therefore, any mapping from $X_{\alpha,\delta}$ into itself that is continuous with respect to the metric d has a fixed point by the Leray–Schauder–Tychonov theorem [Ree80, Thm. V.19].

Note further that each $X_{\alpha,\delta}$ contains a maximal element with respect to the partial order $\boldsymbol{a} \leq \boldsymbol{b}$ defined by $a_k \leq b_k$ for all $k \in \mathbb{N}_0$, which is given by $(1, \alpha, \delta^2, \delta^3, \ldots)$. This property finally leads to

Proposition 7. *The space $P_{\alpha,\delta} = \{\boldsymbol{p} \in \mathcal{M}_1^+ : \boldsymbol{a}(\boldsymbol{p}) \in X_{\alpha,\delta}\}$, equipped with (the metric induced by) the total variation norm, is compact and convex.*

The proof is based on the following two lemmas.

Lemma 4. *For any subset of $P_{\alpha,\delta}$, the corresponding generating functions from (26) are locally bounded on $B_{1+1/\delta}(0)$.*

Proof: It is sufficient to show boundedness on every compact $K \subset B_{1+1/\delta}(0)$, see [Rem98, Sec. 7.1]. Thus, let such a K be given and fix $r \in [0, \frac{1}{\delta}[$ such that $K \subset \overline{B_{1+r}(0)}$. Then, for every $\boldsymbol{p} \in P_{\alpha,\delta}$ and every $z \in K$,

$$\begin{aligned} |\psi(z)| = \Big|\sum_{k\geq 0} p_k z^k\Big| \leq \sum_{k\geq 0} p_k (1+r)^k = \psi(1+r) &= \sum_{k\geq 0} (k+1)\boldsymbol{a}(\boldsymbol{p})_k \, r^k \\ &\leq 1 + 2\,\alpha\, r + \sum_{k\geq 2} (k+1)(r\delta)^k < \infty\,, \end{aligned} \tag{34}$$

where $r\delta < 1$ was used. This needed to be shown. □

Lemma 5. *If for a sequence $(\boldsymbol{p}^{(n)}) \subset P_{\alpha,\delta}$ the coefficients $\boldsymbol{a}^{(n)} = \boldsymbol{a}(\boldsymbol{p}^{(n)})$ from (29) converge to some $\boldsymbol{a}$ with respect to the metric d from (31), then the generating functions ψ_n from (26) converge compactly to some ψ with $\psi(z) = \sum_{k\geq 0} p_k z^k$ and thus the $\boldsymbol{p}^{(n)}$ converge in norm to $\boldsymbol{p} \in P_{\alpha,\delta}$.*

Proof: According to Lemma 4, the sequence (ψ_n) is locally bounded in $B_{1+1/\delta}(0)$. Due to the pointwise convergence $|a_k^{(n)} - a_k| \leq d_k^{-1} d(\boldsymbol{a}^{(n)}, \boldsymbol{a}) \to 0$, we have

$$\psi_n^{(k)}(1) = (k+1)!\, a_k^{(n)} \xrightarrow{n\to\infty} (k+1)!\, a_k = \psi^{(k)}(1) \tag{35}$$

for every $k \in \mathbb{N}_0$. Then, the compact convergence $\psi_n \to \psi$ follows from Vitali's theorem [Rem98, Thm. 7.3.2]. In particular, this implies the convergence $p_k^{(n)} \to p_k \geq 0$ and $1 = \sum_{k\geq 0} p_k^{(n)} = \psi_n(1) \to \psi(1) = \sum_{k\geq 0} p_k$, thus $\boldsymbol{p} \in \mathcal{M}_1^+$.

Now choose $r \in]1, 1 + \frac{1}{\delta}[$. For every $\varepsilon > 0$, there is an n_ε such that, for all $n \geq n_\varepsilon$, $\sup_{|z|\leq r} |\psi(z) - \psi_n(z)| < \varepsilon$. This implies

$$|p_k^{(n)} - p_k| = \left| \frac{1}{2\pi i} \oint_{|z|=r} \frac{\psi_n(z) - \psi(z)}{z^{k+1}} \,\mathrm{d}z \right| < \frac{\varepsilon}{r^k} \tag{36}$$

for all $n \geq n_\varepsilon$ by Cauchy's integral formula [Lan99, Thm. 7.3]. Now, let $\varepsilon > 0$ be given. Then

$$\|\boldsymbol{p}^{(n)} - \boldsymbol{p}\|_1 = \sum_{k \geq 0} |p_k^{(n)} - p_k| < \varepsilon \frac{1}{1 - \frac{1}{r}} \tag{37}$$

for all $n \geq n_\varepsilon$, which proves the claim. □

Proof of Proposition 7: Let $(\boldsymbol{p}^{(n)})$ denote some (arbitrary) sequence in $P_{\alpha,\delta}$ and $(\boldsymbol{a}^{(n)}) = (\boldsymbol{a}(\boldsymbol{p}^{(n)}))$ the corresponding sequence in $X_{\alpha,\delta}$. Due to Proposition 5, there is a convergent subsequence $(\boldsymbol{a}^{(n_i)})_i$. Then, by Lemma 5, $(\boldsymbol{p}^{(n_i)})_i$ converges in norm to some $\boldsymbol{p} \in P_{\alpha,\delta}$. This proves the compactness property. The convexity of $P_{\alpha,\delta}$ is a simple consequence of the convexity of $\mathcal{M}_1^+$, the linearity of the mapping $\boldsymbol{a}$, and the convexity of $X_{\alpha,\delta}$. □

Another property of the mapping $\boldsymbol{a}\colon P_{\alpha,\delta} \to X_{\alpha,\delta}$ is stated in

Lemma 6. *For every α and δ, the mapping $\boldsymbol{a}\colon P_{\alpha,\delta} \to X_{\alpha,\delta}$ from* (29) *is continuous (with respect to the total variation norm and the metric d) and injective. Its inverse $\boldsymbol{p}\colon \boldsymbol{a}(P_{\alpha,\delta}) \to P_{\alpha,\delta}$ is continuous as well.*

Proof: Let $\boldsymbol{p}$, $\boldsymbol{q} \in P_{\alpha,\delta}$ and assume $\boldsymbol{a}(\boldsymbol{p}) = \boldsymbol{a}(\boldsymbol{q})$. Then, as in the proof of Lemma 3, the uniqueness of the generating function in $B_{1+1/\delta}(0)$ follows, and thus $\boldsymbol{p} = \boldsymbol{q}$, which proves the injectivity of $\boldsymbol{a}$. The other statements follow from Vitali's theorem [Rem98, Thm. 7.3.2]: Norm convergence of a sequence $(\boldsymbol{p}^{(n)}) \subset P_{\alpha,\delta}$ to some $\boldsymbol{p}$ implies convergence of its element sequences and thus compact convergence of the corresponding generating functions ψ_n to ψ, where $\psi(z) = \sum_{k \geq 0} p_k\, z^k$. This, in turn, implies convergence of each sequence $(\boldsymbol{a}(\boldsymbol{p}^{(n)})_k)$ to $\boldsymbol{a}(\boldsymbol{p})_k$, from which, as in (33), the convergence $(\boldsymbol{a}(\boldsymbol{p}^{(n)})) \to \boldsymbol{a}(\boldsymbol{p})$ (with respect to d) follows. The converse is the statement of Lemma 5 (see also [Ped89, Prop. 1.6.8]). □

Note that, if $\rho(\psi) > 2$, the inverse of the mapping $\boldsymbol{a}$ is given by

$$\boldsymbol{p}(\boldsymbol{a})_k = \sum_{\ell \geq k} (-1)^{\ell - k} \binom{\ell}{k} (\ell + 1)\, a_\ell \,.$$

5 Random unequal crossover

Let us now turn to the random UC model, described by $q = 1$ in (8). Here, the recombinator (2) simplifies to [Shp02, (3.1)]

$$\mathcal{R}_1(\boldsymbol{p})_i = \sum_{\substack{k,\ell \geq 0 \\ k+\ell \geq i}} \frac{1 + \min\{k, \ell, i, k + \ell - i\}}{(k+1)(\ell+1)}\, p_k\, p_\ell \,. \tag{38}$$

As for internal UC, by Lemma 1, the reversibility condition (17) directly leads to an expression for fixed points,

$$\frac{p_k}{k+1} \frac{p_\ell}{\ell+1} = \frac{p_i}{i+1} \frac{p_j}{j+1} \quad \text{for all } k + \ell = i + j \,. \tag{39}$$

This has $p_k = C(k+1)x^k$ as a solution, with appropriate parameter x and normalization constant C. Again, it turns out that all fixed points are given this way, as stated by

Proposition 8 [Shp02, Thm. A.2]. *Every fixed point $\boldsymbol{p} \in \mathcal{M}_1^+$ of $\mathcal{R}_1$ is of the form*

$$p_k = \left(\frac{2}{m+2}\right)^2 (k+1) \left(\frac{m}{m+2}\right)^k , \tag{40}$$

where $m = \sum_{k \geq 0} k\, p_k \geq 0$.

One can verify this in several ways, one being a direct inductive calculation.

The main result of this section is

Theorem 3. *Assume that $\limsup_{k\to\infty} \sqrt[k]{p_k(0)} < 1$. Then, both in discrete and in continuous time, $\lim_{t\to\infty} \|\boldsymbol{p}(t) - \boldsymbol{p}\|_1 = 0$ with the appropriate fixed point $\boldsymbol{p}$ from Proposition 8.*

For a proof, we consider the following alternative process, verbally described in [Shp02, p. 720f], which induces the same dynamics as random UC. This ultimately leads to a simple expression for the induced recombinator of the coefficients $\boldsymbol{a}$ from (29), which allows for an explicit solution.

Proposition 9. *Let $\boldsymbol{p} \in \mathcal{M}_1^+$. Then,*

$$\pi_k = \sum_{\ell \geq k} \frac{1}{\ell+1} p_\ell \tag{41}$$

gives a probability measure $\boldsymbol{\pi} \in \mathcal{M}_1^+$, and the recombinator can be written as

$$\mathcal{R}_1(\boldsymbol{p})_i = \sum_{j=0}^{i} \pi_j\, \pi_{i-j} = (\boldsymbol{\pi} * \boldsymbol{\pi})_i\, , \tag{42}$$

where $*$ denotes the convolution in $\ell^1(\mathbb{N}_0)$.

Proof: It is easily verified that $\boldsymbol{\pi}$ is normalized to 1. With respect to (42), note the following identity for $k + \ell \geq i$,

$$|\{j : (i-\ell) \vee 0 \leq j \leq i \wedge k\}| = 1 + \min\{k, \ell, i, k+\ell-i\}\, , \tag{43}$$

which can be shown by treating the four cases in the LHS separately. With this, inserting (41) into the RHS of (42) yields

$$\begin{aligned} \sum_{j=0}^{i} \pi_j\, \pi_{i-j} &= \sum_{j=0}^{i} \sum_{k \geq j} \sum_{\ell \geq i-j} \frac{1}{(k+1)(\ell+1)}\, p_k\, p_\ell \\ &= \sum_{\substack{k,\ell \geq 0 \\ k+\ell \geq i}} \frac{1}{(k+1)(\ell+1)}\, p_k\, p_\ell \sum_{j=(i-\ell)\vee 0}^{i \wedge k} 1 \\ &= \sum_{\substack{k,\ell \geq 0 \\ k+\ell \geq i}} \frac{1 + \min\{k, \ell, i, k+\ell-i\}}{(k+1)(\ell+1)}\, p_k\, p_\ell \;=\; \mathcal{R}_1(\boldsymbol{p})_i\, . \end{aligned} \tag{44}$$

□

Here, (41) describes a process in which, without any pairing, each sequence is cut equally likely between two of its building blocks. In a second step, described by (42), these fragments are paired randomly and joined.

This nice structure has an analog on the level of the generating functions.

Proposition 10. *Let $\phi(z) = \sum_{k\geq 0} \pi_k z^k$ denote the generating function for $\boldsymbol{\pi}$ from* (41)*. Then*

$$\phi(z) = \frac{1}{1-z}\int_z^1 \psi(\zeta)\,\mathrm{d}\zeta \qquad \text{and} \qquad \mathcal{R}_1(\psi)(z) = \phi(z)^2\,. \tag{45}$$

Proof: Equations (41) and (42) lead to

$$\begin{aligned} \phi(z) &= \sum_{k\geq 0}\sum_{\ell\geq k} \frac{1}{\ell+1} p_\ell z^k = \sum_{\ell\geq 0}\frac{1}{\ell+1}p_\ell \sum_{k\leq \ell} z^k = \sum_{\ell\geq 0}\frac{1}{\ell+1}p_\ell \frac{1-z^{\ell+1}}{1-z} \\ &= \frac{1}{1-z}\sum_{\ell\geq 0} p_\ell \frac{1-z^{\ell+1}}{\ell+1} = \frac{1}{1-z}\int_z^1 \psi(\zeta)\,\mathrm{d}\zeta \end{aligned} \tag{46}$$

and, due to absolute convergence of the series involved,

$$\mathcal{R}_1(\psi)(z) = \sum_{k\geq 0}\mathcal{R}_1(\boldsymbol{p})_k z^k = \sum_{k\geq 0} z^k \sum_{\ell=0}^{k}\pi_\ell \pi_{k-\ell} = \sum_{\ell\geq 0}\pi_\ell z^\ell \sum_{k\geq \ell}\pi_{k-\ell} z^{k-\ell} = \phi(z)^2\,. \tag{47}$$

□

The following lemma states that the radius of convergence of ψ does not decrease under the random UC dynamics. Thus, it is ensured that, if $\rho(\psi) > 1$, also $\mathcal{R}_1(\psi)$ may be described by an expansion at $z = 1$, i.e., by coefficients $\boldsymbol{a}$.

Lemma 7. *The radius of convergence of $\mathcal{R}_1(\psi)$ is $\rho(\mathcal{R}_1(\psi)) \geq \rho(\psi)$.*

Proof: As $1/\rho(\psi) = \limsup_{k\to\infty} \sqrt[k]{p_k} =: x \leq 1$ and $\lim_{k\to\infty}\sqrt[k]{k+1} = 1$, there is a constant $C > 0$ with $p_k \leq C(k+1)x^k$ for all k. Note the identity

$$\sum_{j=0}^{n}(1+\min\{i,j,n-i,n-j\}) = (i+1)(n-i+1) \tag{48}$$

for $i \leq n$, which follows from an elementary calculation. Then (38) implies

$$\begin{aligned} \mathcal{R}_1(\boldsymbol{p})_i &\leq C^2 \sum_{\substack{k,\ell\geq 0\\ k+\ell\geq i}} x^{k+\ell}(1+\min\{k,\ell,i,k+\ell-i\}) \\ &= C^2\sum_{n\geq i} x^n \sum_{j=0}^{n}(1+\min\{i,j,n-i,n-j\}) \\ &= C^2(i+1)x^i\sum_{\ell\geq 0}(\ell+1)\,x^\ell = \left(\frac{C}{1-x}\right)^2 (i+1)\,x^i\,. \end{aligned} \tag{49}$$

Accordingly, $\limsup_{k\to\infty}\sqrt[k]{\mathcal{R}_1(\boldsymbol{p})_k} \leq x \leq 1$, which proves the claim. □

These results enable us to derive the following expression for the coefficients $\boldsymbol{a}$, using the expansion of (28):

$$\mathcal{R}_1(\psi)(z) = \left[\frac{1}{z-1}\int_1^z \psi(\zeta)\,\mathrm{d}\zeta\right]^2 = \left[\sum_{k\geq 0} a_k (z-1)^k\right]^2 = \sum_{k\geq 0}\left(\sum_{n=0}^{k} a_n a_{k-n}\right)(z-1)^k\,. \tag{50}$$

So it is natural to define the induced recombinator

$$\tilde{\mathcal{R}}_1(\boldsymbol{a})_k = \frac{1}{k+1}\sum_{n=0}^{k} a_n a_{k-n} \geq 0\,, \tag{51}$$

for which we have

Lemma 8. *The recombinator $\tilde{\mathcal{R}}_1$ given by* (51) *maps each space $X_{\alpha,\delta}$ into itself and is continuous with respect to the metric d from* (31).

Proof: Let α, $\delta > 0$ be given and $\boldsymbol{a}$, $\boldsymbol{b} \in X_{\alpha,\delta}$. Trivially, $\tilde{\mathcal{R}}_1(\boldsymbol{a})_0 = 1$ and $\tilde{\mathcal{R}}_1(\boldsymbol{a})_1 = \alpha$. For $k \geq 2$, $\tilde{\mathcal{R}}_1(\boldsymbol{a})_k = \frac{1}{k+1}\sum_{\ell\leq k} a_\ell\, a_{k-\ell} \leq \delta^k$. This proves the first statement. For the continuity, note first that every $\tilde{\mathcal{R}}_1(\boldsymbol{a})_k$ with $k \geq 2$ is continuous as a mapping from $X_{\alpha,\delta}$ to $[0,\delta^k]$. Now, let $\varepsilon > 0$ be given. Choose n large enough so that $\sum_{k>n}(2\gamma)^k < \varepsilon/2$, where γ is the parameter introduced in Definition 3. Then, there is an $\eta > 0$ such that $\sum_{k=2}^{n}(\gamma/\delta)^k|\tilde{\mathcal{R}}_1(\boldsymbol{a})_k - \tilde{\mathcal{R}}_1(\boldsymbol{b})_k| < \varepsilon/2$ for $\boldsymbol{a}$, $\boldsymbol{b} \in X_{\alpha,\delta}$ with $d(\boldsymbol{a},\boldsymbol{b}) < \eta$. Thus, for such $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$d(\tilde{\mathcal{R}}_1(\boldsymbol{a}),\tilde{\mathcal{R}}_1(\boldsymbol{b})) \leq \sum_{k=0}^{n}\left(\frac{\gamma}{\delta}\right)^k|\tilde{\mathcal{R}}_1(\boldsymbol{a})_k - \tilde{\mathcal{R}}_1(\boldsymbol{b})_k| + \sum_{k>n}(2\gamma)^k < \varepsilon\,, \tag{52}$$

which proves the claim. □

Note that the fixed point equation on the level of the coefficients $\boldsymbol{a}$ is always satisfied for a_0 and a_1. If $k > 1$, one obtains the recursion

$$a_k = \frac{1}{k-1}\sum_{n=1}^{k-1} a_n a_{k-n}\,, \tag{53}$$

which shows that at most one fixed point with given mean can exist.

Let us now consider the discrete time case first. Analogously to (6), define the coefficients belonging to $\boldsymbol{p}(t)$ as $\boldsymbol{a}(t) = \boldsymbol{a}(\boldsymbol{p}(t))$, which are assumed to exist. It is clear from (27), (50), and (51) that $\boldsymbol{a}(t+1) = \tilde{\mathcal{R}}_1(\boldsymbol{a}(t))$. We then have the following two propositions.

Proposition 11. *Assume $\boldsymbol{a}(0)$ to exist. Then, in discrete time,* $\lim_{t\to\infty} a_k(t) = \alpha^k$, *for all $k \geq 0$.*

This result indicates that a weaker condition than the one of Theorem 3 may be sufficient for convergence of $\boldsymbol{p}(t)$.

PROOF: Clearly, $a_0(t) \equiv 1$, $a_1(t) \equiv \alpha$. Furthermore, by the assumption and (50), the coefficients $a_k(t)$ exist for all k, $t \in \mathbb{N}_0$. Now, assume that the claim holds for all $k \leq n$ with some n and let $k = n+1$. According to the properties of lim sup and lim inf, we have

$$\limsup_{t\to\infty} a_k(t+1) \leq \frac{1}{k+1} \sum_{\ell=0}^{k} \limsup_{t\to\infty} \big(a_\ell(t) a_{k-\ell}(t)\big) = \frac{k-1}{k+1}\alpha^k + \frac{2}{k+1} \limsup_{t\to\infty} a_k(t) \quad (54)$$

and analogously with $\geq$ for lim inf. This leads to

$$\frac{k-1}{k+1} \limsup_{t\to\infty} a_k(t) \leq \frac{k-1}{k+1}\alpha^k \leq \frac{k-1}{k+1} \liminf_{t\to\infty} a_k(t)\,, \quad (55)$$

from which the claim follows for all $k \leq n+1$ and, by induction over n, for all $k \geq 0$. □

Proposition 12. *The recombinator $\tilde{\mathcal{R}}_1$, acting on $X_{\alpha,\delta}$, is a strict contraction with respect to the metric d from* (31)*, i.e., there is a $C < 1$ such that*

$$d(\tilde{\mathcal{R}}_1(\boldsymbol{a}), \tilde{\mathcal{R}}_1(\boldsymbol{b})) \leq C\, d(\boldsymbol{a}, \boldsymbol{b}) \quad (56)$$

for all $\boldsymbol{a}$, $\boldsymbol{b} \in X_{\alpha,\delta}$.

PROOF: First consider, for $k \geq 2$, without using the special choice of the d_k,

$$\begin{aligned} d(\tilde{\mathcal{R}}_1(\boldsymbol{a}), \tilde{\mathcal{R}}_1(\boldsymbol{b})) &= \sum_{k\geq 2} d_k \frac{1}{k+1} \Big| \sum_{\ell=0}^{k} (a_\ell\, a_{k-\ell} - b_\ell\, b_{k-\ell}) \Big| \\ &= \sum_{k\geq 2} d_k \frac{1}{k+1} \Big| \sum_{\ell=0}^{k} (a_\ell - b_\ell)(a_{k-\ell} + b_{k-\ell}) \Big| \quad (57) \\ &\leq \sum_{k\geq 2} d_k \frac{2}{k+1} \sum_{\ell=2}^{k} \delta^{k-\ell} |a_\ell - b_\ell| = \sum_{\ell\geq 2} d_\ell |a_\ell - b_\ell| \sum_{k\geq\ell} \frac{2}{k+1} \delta^{k-\ell} \frac{d_k}{d_\ell}\,. \end{aligned}$$

With the choice $d_k = (\gamma/\delta)^k$, where we had $\gamma < \frac{1}{3}$, we can find, for $\ell \geq 2$, an upper bound for the inner sum,

$$\sum_{k\geq\ell} \frac{2}{k+1} \delta^{k-\ell} \frac{d_k}{d_\ell} \leq \frac{2}{3} \sum_{k\geq\ell} \gamma^{k-\ell} = \frac{2}{3-3\gamma} =: C < 1\,, \quad (58)$$

which, together with (57), proves the claim. □

Together with Banach's fixed point theorem (compare [Ree80, Thm. V.18]), the two propositions imply that $\boldsymbol{a}(t)$ converges to $(1, \alpha, \alpha^2, \ldots)$ with respect to the metric d, and that convergence is exponentially fast.

In continuous time, we consider the time derivative of $\boldsymbol{a}(t) := \boldsymbol{a}(\boldsymbol{p}(t))$, which is, by (29),

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a}(t) = \tfrac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{a}\big(\boldsymbol{p}(t)\big) = \boldsymbol{a}\big(\mathcal{R}_1(\boldsymbol{p}(t)) - \boldsymbol{p}(t)\big) = \tilde{\mathcal{R}}_1(\boldsymbol{a}(t)) - \boldsymbol{a}(t)\,. \quad (59)$$

The following lemma ensures, together with [Ama90, Thm. 7.6 and Rem. 7.10(b)], that this initial value problem has a unique solution for all $\boldsymbol{a}(0) = \boldsymbol{a}_0 \in X_{\alpha,\delta}$.

Lemma 9. *Consider the Banach space $H_{\gamma/\delta}$ from* (32), *with some $0 < \gamma < \frac{1}{3}$, and its open subset $Y = \{\boldsymbol{x} \in H_{\gamma/\delta} : |x_k| < (2\delta)^k\}$. Then, the recombinator $\tilde{\mathcal{R}}_1$ from* (51) *maps Y into itself, satisfies a global Lipschitz condition, and is bounded on Y. Furthermore, it is infinitely differentiable, $\tilde{\mathcal{R}}_1 \in \mathrm{C}^\infty(Y, Y)$.*

Proof: For $\boldsymbol{x} \in Y$, one has $|x_k| < (2\delta)^k$, hence $|\tilde{\mathcal{R}}_1(\boldsymbol{x})_k| < (2\delta)^k$ with a similar argument as in the proof of Lemma 8. Consequently, $\tilde{\mathcal{R}}_1(Y) \subset Y$. So let $\boldsymbol{x}, \boldsymbol{y} \in Y$. Then, similarly to the proof of Proposition 12, one shows the Lipschitz condition

$$\|\tilde{\mathcal{R}}_1(\boldsymbol{x}) - \tilde{\mathcal{R}}_1(\boldsymbol{y})\| \;\le\; \sum_{\ell\ge 0} \left(\frac{\gamma}{\delta}\right)^\ell |x_\ell - y_\ell| \sum_{k\ge\ell} \frac{2}{k+1}(2\gamma)^{k-\ell} \;\le\; \frac{2}{1-2\gamma}\|\boldsymbol{x} - \boldsymbol{y}\| \tag{60}$$

and, since $\|\boldsymbol{x}\| < 1/(1-2\gamma)$ in Y, the boundedness,

$$\|\tilde{\mathcal{R}}_1(\boldsymbol{x})\| \le \frac{1}{1-2\gamma}\|\boldsymbol{x}\| < \frac{1}{(1-2\gamma)^2}\,. \tag{61}$$

With respect to differentiability, consider, for sufficiently small $\boldsymbol{h} \in Y$,

$$\tilde{\mathcal{R}}_1(\boldsymbol{x}+\boldsymbol{h})_k = \tilde{\mathcal{R}}_1(\boldsymbol{x})_k + \frac{2}{k+1}\sum_{\ell=0}^{k} x_{k-\ell}\, h_\ell + \tilde{\mathcal{R}}_1(\boldsymbol{h})_k\,. \tag{62}$$

Since

$$\|\tilde{\mathcal{R}}_1(\boldsymbol{h})\| \le \sum_{k\ge 0}\left(\frac{\gamma}{\delta}\right)^k \frac{1}{k+1}\sum_{\ell=0}^{k}|h_{k-\ell}|\,|h_\ell| = \sum_{\ell\ge 0}\left(\frac{\gamma}{\delta}\right)^\ell |h_\ell| \sum_{k\ge\ell}\left(\frac{\gamma}{\delta}\right)^{k-\ell}\frac{|h_{k-\ell}|}{k+1} \le \|\boldsymbol{h}\|^2\,, \tag{63}$$

it is clear that $\tilde{\mathcal{R}}_1$ is differentiable with linear (and thus continuous) derivative, whose Jacobi matrix is explicitly $\tilde{\mathcal{R}}_1'(\boldsymbol{x})_{k\ell} = \frac{\partial}{\partial x_\ell}\tilde{\mathcal{R}}_1(\boldsymbol{x})_k = \frac{2}{k+1}x_{k-\ell}$ if $k \ge \ell$ and zero otherwise, so $\tilde{\mathcal{R}}_1 \in \mathrm{C}^1(Y, Y)$. It is now trivial to show that $\tilde{\mathcal{R}}_1 \in \mathrm{C}^2(Y, Y)$ with constant second derivative and thus $\tilde{\mathcal{R}}_1 \in \mathrm{C}^\infty(Y, Y)$. □

Proposition 13. *If $\boldsymbol{a}_0 \in X_{\alpha,\delta}$ for some α, δ, then $\boldsymbol{a}(t) \in X_{\alpha,\delta}$ for all $t \ge 0$ and $\lim_{t\to\infty} d(\boldsymbol{a}(t), \boldsymbol{\alpha}) = 0$ with $\boldsymbol{\alpha} = (1, \alpha, \alpha^2, \alpha^3, \ldots)$.*

Proof: The first statement follows from [Mar76, Thm. VI.2.1] (see also [Ama90, Thm. 16.5]) since, due to the convexity of $X_{\alpha,\delta}$, we have $\boldsymbol{a}+t(\tilde{\mathcal{R}}(\boldsymbol{a})-\boldsymbol{a}) \in X_{\alpha,\delta}$ for every $\boldsymbol{a} \in X_{\alpha,\delta}$ and $t \in [0,1]$, hence a subtangent condition is satisfied. For the second statement, observe that $\tilde{\mathcal{R}}_1(\boldsymbol{\alpha}) = \boldsymbol{\alpha}$. We now show that

$$L(\boldsymbol{a}_0) = d(\boldsymbol{a}_0, \boldsymbol{\alpha}) \tag{64}$$

is a Lyapunov function, cf. Definition 1. With the notation of Lemma 9, note that the compact metric space $X_{\alpha,\delta}$ is contained in the open subset Y of the Banach space $H_{\gamma/\delta}$. The continuity of L is obvious. Now, let $\boldsymbol{a}_0 \in X_{\alpha,\delta}$ be given. By Lemma 9 and [Ama90, Thm. 9.5 and Rem. 9.6(b)], the solution $\boldsymbol{a}(t)$ of (59) is infinitely differentiable. Thus, for $t \in [0,1]$,

$$\begin{aligned} L(\boldsymbol{a}(t)) - L(\boldsymbol{a}_0) &= \|\boldsymbol{a}_0 + t(\tilde{\mathcal{R}}_1(\boldsymbol{a}_0) - \boldsymbol{a}_0) + o(t) - \boldsymbol{\alpha}\| - \|\boldsymbol{a}_0 - \boldsymbol{\alpha}\| \\ &\leq t\big(\|\tilde{\mathcal{R}}_1(\boldsymbol{a}_0) - \tilde{\mathcal{R}}_1(\boldsymbol{\alpha})\| - \|\boldsymbol{a}_0 - \boldsymbol{\alpha}\|\big) + o(t)\,, \end{aligned} \tag{65}$$

where $o(t)$ is the usual Landau symbol and represents some function that vanishes faster than t as $t \to 0$. From this, by the strict contraction property of $\tilde{\mathcal{R}}_1$ (Proposition 12), the Lyapunov property (13) follows, with equality if and only if $\boldsymbol{a}_0 = \boldsymbol{\alpha}$. Since $X_{\alpha,\delta}$ is compact, the claim follows from Theorem 1. □

We are now able to give the previously postponed

Proof of Theorem 3: It follows from Proposition 6 that $\boldsymbol{a}(0) = \boldsymbol{a}(\boldsymbol{p}(0)) \in X_{\alpha,\delta}$ with $\alpha = \frac{1}{2}m$ and some δ. In discrete time, according to Propositions 11 and 12 and Banach's fixed point theorem (compare [Ree80, Thm. V.18]), $\boldsymbol{a}(t) \to \boldsymbol{\alpha} = (1, \alpha, \alpha^2, \ldots)$ with respect to the metric d. Letting $x = m/(m+2)$ and inserting (40) into (29) yields

$$a_k = \sum_{\ell \geq k} \frac{\ell!}{(\ell-k)!(k+1)!}(1-x)^2(\ell+1)x^\ell = (1-x)^2 \sum_{\ell \geq k} \binom{\ell+1}{k+1} x^\ell = \left(\frac{x}{1-x}\right)^k = \alpha^k\,. \tag{66}$$

The claim now follows from Lemma 5. Similarly, in continuous time, the claim follows from Proposition 13. □

Let us finally note

Proposition 14. *For the dynamics described by* (59)*, the fixed point* $\boldsymbol{\alpha}$ *from Proposition* 13 *is exponentially stable.*

Proof: Let $\boldsymbol{a}_0 \in X_{\alpha,\delta}$ be arbitrary. The Lyapunov function from the proof of Proposition 13 satisfies, as a consequence of (65) and Proposition 12,

$$\dot{L}(\boldsymbol{a}_0) \leq d(\tilde{\mathcal{R}}_1(\boldsymbol{a}_0), \tilde{\mathcal{R}}_1(\boldsymbol{\alpha})) - d(\boldsymbol{a}_0, \boldsymbol{\alpha}) \leq -(1-C)\, d(\boldsymbol{a}_0, \boldsymbol{\alpha})\,, \tag{67}$$

with $0 < C < 1$. From this, together with (64) and [Ama90, Thm. 18.7], the claim follows. □

Furthermore, in a UC model introduced by Takahata [Tak81], for which

$$T_{ij,k\ell} = \delta_{i+j,k+\ell} \frac{1}{k+\ell+1}\,,$$

the recombinator $\tilde{\mathcal{R}}_1$ appears for the coefficients $\boldsymbol{b}(\boldsymbol{p})_k = (k+1)\, \boldsymbol{a}(\boldsymbol{p})_k$, where $\boldsymbol{b}(\boldsymbol{p})_1$ is the mean copy number m. The above results then imply, under the appropriate condition on $\boldsymbol{p}(0)$, that $\boldsymbol{b}(t) \to (1, m, m^2, \ldots)$ as $t \to \infty$ both in discrete and in continuous time. This corresponds to convergence of $\boldsymbol{p}(t)$ to the fixed point $\boldsymbol{p}$ with $p_k = \frac{1}{m+1}(\frac{m}{m+1})^k$.

6 The intermediate parameter regime

In this section, q may take any value in $[0,1]$. With respect to reversibility of fixed points, one finds

Proposition 15. *For intermediate parameter values $q \in \,]0,1[$, any fixed point $\boldsymbol{p} \in \mathcal{M}_1^+$ of the recombinator $\mathcal{R}_q$, given by (2) and (8), satisfies $p_k > 0$ for all $k \geq 0$ (unless it is the trivial fixed point $\boldsymbol{p} = (1,0,0,\ldots)$ we excluded). None of these extra fixed points is reversible.*

PROOF: Let a non-trivial fixed point $\boldsymbol{p}$ be given and choose any $n > 0$ with $p_n > 0$. Observe that $T^{(q)}_{n+1\,n-1,nn} > 0$ for $0 < q < 1$ and hence

$$p_{n\pm1} = \mathcal{R}_q(\boldsymbol{p})_{n\pm1} = \sum_{j,k,\ell\geq0} T^{(q)}_{n\pm1\,j,k\ell}\, p_k\, p_\ell \geq T^{(q)}_{n+1\,n-1,nn}\, p_n\, p_n > 0\,. \tag{68}$$

The first statement follows by induction. For the second statement, evaluate the reversibility condition (17) for all combinations of i, j, k, ℓ with $i + j = k + \ell \leq 4$. This leads to four independent equations. Three of them can be transformed to the recursion

$$p_k = \frac{(k+1)q}{2(k-1)+2q}\,\frac{p_1}{p_0}\,p_{k-1}\,, \qquad k \in \{2,3,4\}\,, \tag{69}$$

from which one derives explicit equations for all p_k with $k \in \{2,3,4\}$ in terms of p_0 and p_1. Inserting the one for p_2 into the remaining equation yields another equation for p_4 in terms of p_0 and p_1, which contradicts the first one for all $q \in \,]0,1[$, as is easily verified. □

So, non-trivial fixed points for $0 < q < 1$ are not reversible, and thus much more difficult to determine. Our most general result so far is

Theorem 4. *If $\boldsymbol{p}(0) \in P_{\alpha,\delta}$ for some α, δ, then $\boldsymbol{p}(t) \in P_{\alpha,\delta}$ for all times $t \in \mathbb{N}_0$, respectively $t \in \mathbb{R}_{\geq 0}$, and $\mathcal{R}_q$ has a fixed point in $P_{\alpha,\delta}$.*

The proof is based on the fact that $\mathcal{R}_q$ is, in a certain sense, monotonic in the parameter q. This is stated in

Proposition 16. *Assume $\boldsymbol{a}(\boldsymbol{p}) \in X_{\alpha,\delta}$ for some α, δ. Then, with respect to the partial order introduced before Proposition 7, $\boldsymbol{a}(\mathcal{R}_q(\boldsymbol{p})) \leq \boldsymbol{a}(\mathcal{R}_{q'}(\boldsymbol{p}))$ for all $0 \leq q \leq q' \leq 1$. In particular, $\boldsymbol{a}(\mathcal{R}_q(\boldsymbol{p})) \in X_{\alpha,\delta}$ for all $0 \leq q \leq 1$.*

To show this, we need three rather technical lemmas. The first one collects formal conditions on the difference of two distributions $T^{(q)}_{ij,k\ell}$ with different parameter values (but $j = k + \ell - i$ and the same fixed k, ℓ). These are then verified in our case.

Lemma 10. *Let the numbers $x_i \in \mathbb{R}$ ($0 \leq i \leq r$ with some $r \in \mathbb{N}_0$) satisfy the following three conditions:*

$$\sum_{i=0}^{r} x_i = 0\,. \tag{70}$$

$$x_{r-i} = x_i \quad \text{for all } 0 \leq i \leq r. \tag{71}$$

$$\text{There is an integer } n \text{ such that } \begin{cases} x_i \geq 0 & \text{for } 0 \leq i \leq n, \\ x_i < 0 & \text{for } n < i \leq \lfloor \frac{r}{2} \rfloor. \end{cases} \tag{72}$$

Further, let $f_i \in \mathbb{R}$ $(0 \le i \le r)$ *be given with*

$$0 \le f_1 - f_0 \le f_2 - f_1 \le \ldots \le f_r - f_{r-1}\,. \tag{73}$$

Then we have

$$\sum_{i=0}^{r} f_i x_i \ge 0\,. \tag{74}$$

PROOF: Let us first consider the trivial cases. If $x_i \equiv 0$, everything is clear, so let $x_i \not\equiv 0$. If $r \le 1$ then $x_i \equiv 0$, so let $r \ge 2$, and thus $n \le \frac{r}{2} - 1$. Define $x_{\frac{r}{2}} = f_{\frac{r}{2}} = 0$ for odd r. Then we can write

$$\sum_{i=0}^{r} f_i x_i = \sum_{i=0}^{n} (f_i + f_{r-i})\, x_i + \sum_{i=n+1}^{\lceil \frac{r}{2} \rceil - 1} (f_i + f_{r-i})\, x_i + f_{\frac{r}{2}} x_{\frac{r}{2}}\,. \tag{75}$$

Further, for $r - i \ge i$, due to (73),

$$f_i + f_{r-i} = f_{i-1} + f_{r-i+1} + (f_i - f_{i-1}) - (f_{r-i+1} - f_{r-i}) \le f_{i-1} + f_{r-i+1}\,. \tag{76}$$

Now, define $C := \sum_{i=0}^{n} x_i = -\sum_{i=n+1}^{\lceil \frac{r}{2} \rceil - 1} x_i - \frac{1}{2} x_{\frac{r}{2}} > 0$, and the claim follows with (75), since $r - n \ge n + 1$ by assumption:

$$\sum_{i=0}^{r} f_i x_i \ge C\,[f_n + f_{r-n} - f_{n+1} - f_{r-n-1}] = C\,[(f_{r-n} - f_{r-n-1}) - (f_{n+1} - f_n)] \ge 0\,. \tag{77}$$

$\square$

Lemma 11. *Let* $j \in \mathbb{N}_0$ *be fixed and* $f_i = (i)_j$, $i \in \mathbb{N}_0$, *where* $(i)_j$ *is the falling factorial, which equals* 1 *for* $j = 0$ *and* $i(i-1)\cdots(i-j+1)$ *for* $j > 0$, *hence* $\frac{i!}{(i-j)!}$ *for* $i \ge j$. *Then condition* (73) *is satisfied.*

PROOF: For $j = 0$, condition (73) is trivially true. Otherwise, each f_i is a polynomial of degree j in i with zeros $\{0, 1, \ldots, j-1\}$, thus $0 = f_1 - f_0 = \ldots = f_{j-1} - f_{j-2}$. Then, for $i \ge j - 1$, the polynomial and all its derivatives are increasing functions since $\lim_{i\to\infty} f_i = \infty$. Therefore, for $i \ge j - 1$, we have $0 \le f_{i+1} - f_i \le f_{i+2} - f_{i+1}$. Hence (73) holds. $\square$

Lemma 12. *For* $0 \le q \le q' \le 1$ *and all* k, ℓ, *letting* $r = k + \ell$ *and* $x_i = T^{(q')}_{ik\ell} - T^{(q)}_{ik\ell}$ *makes* (70)–(72) *true (where* $T^{(q)}_{ik\ell} = T^{(q)}_{ij,k\ell}$ *with* $j = k + \ell - i$*).*

PROOF: The validity of (70) and (71) is clear from the normalization (3) and the symmetry of the $T^{(q)}_{ik\ell}$. For (72), let $k \le \ell$ without loss of generality. In the trivial cases $q = q'$ or $k = 0$, choose $n = \lfloor \frac{r}{2} \rfloor$. Otherwise, observe that $x_i = T^{(q')}_{ik\ell} - T^{(q)}_{ik\ell} < 0$ for $k \le i \le \lfloor \frac{r}{2} \rfloor$, since $C^{(q')}_{k\ell} < C^{(q)}_{k\ell}$, and $x_0 > 0$. For $0 \le i \le k$, consider

$$y_i = \frac{x_i}{T^{(q)}_{ik\ell}} + 1 = \frac{C^{(q')}_{k\ell}}{C^{(q)}_{k\ell}} \left(\frac{q'}{q}\right)^{k-i}\,. \tag{78}$$

Here, the first factor is less than 1, the second is equal to 1 for $k = i$, greater than 1 for $0 \le k < i$, and strictly decreasing with i. Since $x_i \ge 0$ if and only if $y_i \ge 1$, there is an index n with the properties needed. $\square$

Proof of Proposition 16: Lemmas 10–12 imply, for all $k,\ \ell,\ j \in \mathbb{N}_0$ with $k+\ell \geq j$,

$$\sum_{i=j}^{k+\ell} \tfrac{i!}{(i-j)!} T^{(q)}_{ik\ell} \leq \sum_{i=j}^{k+\ell} \tfrac{i!}{(i-j)!} T^{(q')}_{ik\ell} \,. \tag{79}$$

Then, since $T^{(q)}_{ik\ell} = 0$ for $i > k+\ell$,

$$\begin{aligned}
\boldsymbol{a}(\mathcal{R}_q(\boldsymbol{p}))_j &= \tfrac{1}{(j+1)!} \sum_{i \geq j} \tfrac{i!}{(i-j)!} \mathcal{R}_q(\boldsymbol{p})_i = \tfrac{1}{(j+1)!} \sum_{i \geq j} \tfrac{i!}{(i-j)!} \sum_{k,\ell \geq 0} T^{(q)}_{ik\ell} p_k p_\ell \\
&= \tfrac{1}{(j+1)!} \sum_{k,\ell \geq 0} p_k p_\ell \sum_{i \geq j} \tfrac{i!}{(i-j)!} T^{(q)}_{ik\ell} \leq \tfrac{1}{(j+1)!} \sum_{k,\ell \geq 0} p_k p_\ell \sum_{i \geq j} \tfrac{i!}{(i-j)!} T^{(q')}_{ik\ell} \\
&= \boldsymbol{a}(\mathcal{R}_{q'}(\boldsymbol{p}))_j \,.
\end{aligned} \tag{80}$$

From this, together with Lemma 8, the claim follows. □

Proof of Theorem 4: According to Proposition 16, $\mathcal{R}_q$ maps $P_{\alpha,\delta}$ into itself, and thus, in discrete time, $\boldsymbol{p}(t) \in P_{\alpha,\delta}$ for every $t \in \mathbb{N}_0$. The analogous statement is true for continuous time $t \in \mathbb{R}_{\geq 0}$. To see this, consider $P_{\alpha,\delta}$ as a closed subset of ℓ^1. Recall that $\mathcal{R}_q - \mathbb{1}$ is globally Lipschitz on ℓ^1 by Proposition 1. Moreover, for any $\boldsymbol{p} \in P_{\alpha,\delta}$ and $t \in [0,1]$, Proposition 7 tells us that

$$\boldsymbol{p} + t(\mathcal{R}_q(\boldsymbol{p}) - \boldsymbol{p}) = (1-t)\boldsymbol{p} + t\mathcal{R}_q(\boldsymbol{p}) \in P_{\alpha,\delta} \,. \tag{81}$$

This implies the positive invariance of $P_{\alpha,\delta}$ by [Mar76, Thm. VI.2.1] (see also [Ama90, Thm. 16.5]). The existence of a fixed point once again follows from the Leray–Schauder–Tychonov theorem [Ree80, Thm. V.19]. □

It is plausible that, given the mean copy number m, never more than one fixed point for $\mathcal{R}_q$ exists. Due to the global convergence results at $q = 0$ and $q = 1$, any non-uniqueness in the vicinity of these parameter values could only come from a bifurcation, not from an independent source. Numerical investigations indicate that no bifurcation is present, but this needs to be analyzed further.

Furthermore, the Lipschitz constant for $\tilde{\mathcal{R}}_q$ can be expected to be continuous in the parameter q, hence to remain strictly less than 1 on the sets $X_{\alpha,\delta}$ in the neighborhood of $q = 1$. So, at least locally, the contraction property should be preserved. Nevertheless, we do not expand on this here since it seems possible to use a rather different approach [Hofa], which has been used for similar problems in game theory, to establish a slightly weaker type of convergence result for all $0 < q < 1$, and probably even on the larger compact set $\mathcal{M}^+_{1,m,C}$ from Lemma 2.

7 Some remarks

We have seen in the preceding sections that, definitely for the extreme cases $q = 0$ and $q = 1$ and presumably for the intermediate values as well, the deterministic dynamics converges to a unique equilibrium solution in both discrete and continuous time. This corresponds to the case of infinite populations. With respect to biological relevance, however, we add some arguments that it is reasonable to expect this to be a good description for large but finite populations as well, i.e., for the underlying (multitype) branching process. For the mutation–selection models of Chapter I, the results by Ethier and Kurtz [Eth86, Thm. 11.2.1] and the generalization [Ath72, Thm. V.7.2] of the Kesten–Stigum theorem [Kes66, Kur97] guaranteed that in the infinite population limit the relative genotype frequencies of the branching process converge almost surely to the deterministic solution (if the population does not go to extinction). These results, however, depend on the finiteness of the genotype space. Since for the UC models considered here the equilibrium distributions are exponentially small for large copy numbers (owing to Theorem 4 also for $q \in\,]0,1[$), one can expect these systems to behave very much like ones with finitely many genotypes. This is also supported by several simulations. Nevertheless, this questions deserves further attention.

Summary and outlook

This thesis was concerned with two model classes of population genetics, one describing the balance of mutation and selection, the other modeling unequal crossover events occuring during recombination. The results will now be summarized and discussed.

With respect to mutation–selection models, the farthest-reaching results were obtained for models with a large but finite set of genotypes. These were presented in Chapter I. After an introduction of the model in its deterministic description (Section 2.1), the underlying branching process was considered (Section 2.2). The equilibrium distribution of the backward process, termed the ancestral distribution, shows up when the matrix governing the forward process, $\boldsymbol{H}$, is symmetrized by a similarity transform to $\tilde{\boldsymbol{H}} = \boldsymbol{S}\boldsymbol{H}\boldsymbol{S}^{-1}$, with a diagonal matrix $\boldsymbol{S}$ (Section 2.3); this is always possible if classes of genotypes sharing the same fitness can be ordered linearly and if mutation only connects neighboring classes, an assumption made for the rest of the chapter. The ancestral distribution determines the response of the mean fitness of the equilibrium population, given by the largest eigenvalue of $\boldsymbol{H}$, to changes in the reproduction rates; it is further connected to a quantity G, termed mutational loss, which is defined as the difference of ancestral and population mean fitness in equilibrium and describes the loss in reproduction rate the population experiences due to mutation (Section 2.5).

The central result is a simple maximum principle (Theorem 1 in Section 3.1) for the mean fitness. Since similarity transforms leave the spectrum invariant, this is also given by the largest eigenvalue of the symmetrized matrix $\tilde{\boldsymbol{H}}$, which can be expressed by Rayleigh's principle. In the limit of an infinite number of mutation classes, in which the set of genotypes densely fills a compact interval, the huge space over which Rayleigh's coefficient is maximized, namely all possible ancestral distributions, effectively reduces to the above-mentioned compact interval of all possible ancestral genotypes in the limit (Section 3.4). Here, the matrix $\boldsymbol{H}$, being the sum of a diagonal matrix $\boldsymbol{R}$ holding the reproduction rates and a Markov generator $\boldsymbol{M}$ describing mutation, are replaced by the fitness function r and a function g connected to the mutational loss G (Proposition 1 and Section 3.5). Technically, for every finite system, upper and lower bounds were derived, which were shown to converge towards each other in the mutation class limit. The same expression directly follows for two further limiting cases (discussed in Section 2.6), the linear case and unidirectional mutation (Sections 3.2 and 3.3). On this basis, explicit expressions for means and variances of fitness and genotype were gained (Theorem 2 and Section 3.6), which are exact in the mutation class limit and in the linear case.

The maximum principle turned out to be an excellent basis to characterize threshold behavior in the equilibrium population when mutation rates are varied relative to the fitness values (Section 4). Since the observation of the so-called error threshold in the quasispecies model with the sharply peaked fitness landscape [Eig71], such behavior has been the object of a controversial debate, leading to a handful of sometimes incompatible definitions. These included a kink in the population mean fitness, the loss of the wildtype from the population, complete mutational degradation, and a jump in the population mean fitness. Now, for the first time in a reasonably large model class, analytical methods could be applied to this problem. The mutation class limit, considered for technical simplification in Sections 2 and 3, turned out to be a necessity for a stringent mathematical definition of the four threshold types just described (Definitions 1–4 in Sections 4.1–4.4), analogously to the thermodynamic limit for the definition of phase transitions in physics. In each case a complete characterization could be given (Theorems 3–6).

It was then a natural next step to ask whether similar results, in particular the derivation of a simple scalar maximum principle, could be achieved for models with a continuous set of genotypes, so-called continuum-of-alleles (COA) models. Some first answers to this question were given in Chapter II. The first investigation was on the connection of COA models to the models with discrete genotypes. Although the limiting genotype set in the mutation class limit of the discrete model class of Chapter I is a continuous interval, there is no well-defined limit model since the mutant distributions become trivial in the limit. However, under some biologically reasonable assumptions, a COA model may be approximated arbitrarily well by models with discrete genotypes, which justifies numerical analyses of continuous models and allows to transfer results. This was shown in Section 2, the main result being Theorem 3 for a compact genotype interval, respectively Theorem 6 for an unbounded interval. Technically, they are generalizations of two standard methods of approximation theory, the Nyström and the Galerkin method, to the COA operators, which are non-compact due to the multiplication part describing reproduction. This problem was solved by considering equivalent compact operators as described in [Bür88, Bür00].

Some first steps towards a simple maximum principle were then taken in Section 3. The necessary ingredients in the discrete case were (i) the possibility to symmetrize the system by a similarity transform with a diagonal matrix, which made the ancestral distribution show up, (ii) the derivation of upper and lower bounds for the mean fitness, and (iii) the existence of a limit in which both bounds converge towards each other. It was rather straightforward to generalize the first point and to characterize all mutation kernels lending themselves to a (global) symmetrization (Proposition 8). The derivation of an upper bound was then very similar to the discrete case (Theorem 7). Establishing a lower bound was technically much more complex, but basically followed the same plan as for discrete genotypes. It was not necessary to require global symmetrizability, though, but an alternative condition of some kind of approximate local symmetrizability sufficed (Proposition 9, Theorem 8). This could not yet be shown to lead to a useful upper bound, which, however, is very plausible to be true if one considers the natural candidate for an analog of the mutation class limit, namely a limit of ever narrower and higher mutant distributions (Section 3.3). As a consequence, intuitively speaking, the ancestral distribution gets sharper and sharper, approaching a delta distribution in the generic case. Thus, local considerations can be expected to be sufficient. But a rigorous proof for a simple maximum principle could, so far, only be given for global symmetrizability and a restricted case (Equation (132)). Numerical comparison indeed corroborated the conjecture that local symmetrizability is sufficient for the existence of a simple maximum principle (Section 3.4).

In this context, recent work by Garske and Grimm [Gar] should be mentioned, who derived a maximum principle for a model with discrete genotypes given as sequences over a four-letter alphabet (representing the four nucleotides). This hints at the general possibility of establishing a simple maximum principle whenever the three ingredients from above are given. It seems furthermore feasible in the near future to treat both discrete and continuous genotype spaces in a unified framework, using the formalism set up in Section 1 of Chapter I.

In Chapter III, a model class for unequal crossover (UC) was treated, which was recently introduced by Shpak and Atteson [Shp02], building on existing models. The authors derived, for two limiting cases, the fixed points of the dynamics, which are uniquely determined by the mean copy number of repeated units under consideration. They further conjectured that any initial distribution, possibly under some mild extra conditions, should converge to the appropriate fixed point—presumably also in the intermediate parameter regime, where alignments with 'overhangs' of the shorter sequence are possible (in contrast to the first limiting case of internal UC) but penalized (as opposed to the second limiting case of random UC). The inherent nonlinearity of the operator describing the UC process, however, makes it a difficult task to prove these conjectures.

In the two limiting cases, this has been possible, for dynamics in both discrete and continuous time, by means of Lyapunov functions and consideration of compact invariant subsets (Theorems 2 and 3). In the intermediate regime, some kind of monotonicity could be shown (Proposition 16), which, unfortunately, is only strong enough to conclude the existence of fixed points (Theorem 4), but not their uniqueness. This remains for future work.

The ultimate aim, of course, is to be able to treat models describing multiple evolutionary factors. To this end, exactly solvable models like those considered in this thesis, which—as a rule of thumb—means that they incorporate two, or at most three processes, may be used as starting points for approximate analyses of additional processes, e.g., by means of perturbation theory or numerics. Thus, in spite of their obvious limitations, they form the basis for a better understanding of biological evolution.

Bibliography

[Ama90] H. Amann. *Ordinary Differential Equations*. De Gruyter, Berlin, 1990.

[Ans71] P. M. Anselone. *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*. Prentice-Hall, Englewood Cliffs, 1971.

[Ath72] K. B. Athreya and P. E. Ney. *Branching Processes*. Springer, New York, 1972.

[Baa] E. Baake and H.-O. Georgii. Multitype branching processes: Interplay of mutation and reproduction. In preparation.

[Baa97] E. Baake, M. Baake, and H. Wagner. Ising quantum chain is equivalent to a model of biological evolution. *Phys. Rev. Lett.* **78** (1997) 559–62. Erratum: *Phys. Rev. Lett.* **79** (1997) 1782.

[Baa98] E. Baake, M. Baake, and H. Wagner. Quantum mechanics versus classical probability in biological evolution. *Phys. Rev.* **E 57** (1998) 1191–1192.

[Baa00] E. Baake and W. Gabriel. Biological evolution through mutation, selection, and drift: An introductory review. In: D. Stauffer (ed.), *Ann. Rev. Comp. Phys.*, vol. 9, (pp. 203–264). World Scientific, 2000.

[Baa01] E. Baake and H. Wagner. Mutation-selection models solved exactly with methods from statistical mechanics. *Genet. Res.* **78** (2001) 93–117.

[Bil99] P. Billingsley. *Convergence of Probability Measures*. 2nd ed. Wiley, New York, 1999.

[Bür88] R. Bürger. Perturbations of positive semigroups and applications to population genetics. *Math. Z.* **197** (1988) 259–272.

[Bür98] R. Bürger. Mathematical properties of mutation–selection models. *Genetica* **102/103** (1998) 279–298.

[Bür00] R. Bürger. *The Mathematical Theory of Selection, Recombination, and Mutation*. Wiley, Chichester, 2000.

[Cha90] B. Charlesworth. Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet. Res. Camb.* **55** (1990) 199–221.

[Cro64] J. F. Crow and M. Kimura. The theory of genetic loads. In: *Proc. XI Int. Congr. Genetics*, vol. 2, (pp. 495–505). Pergamon Press, Oxford, 1964.

[Cro70] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. Harper & Row, New York, 1970.

[Dem85] L. Demetrius, P. Schuster, and K. Sigmund. Polynucleotide evolution and branching processes. *Bull. Math. Biol.* **47** (1985) 239–262.

[Eig71] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwiss.* **58** (1971) 465–523.

[Eig89] M. Eigen, J. S. McCaskill, and P. Schuster. The molecular quasi-species. *Adv. Chem. Phys.* **75** (1989) 149–263.

[Eng97] H. W. Engl. *Integralgleichungen.* Springer, Wien, 1997.

[Eth86] S. N. Ethier and T. G. Kurtz. *Markov Processes – Characterization and Convergence.* Wiley, New York, 1986.

[Ewe79] W. Ewens. *Mathematical Population Genetics.* Springer, Berlin, 1979.

[Fra97] S. Franz and L. Peliti. Error threshold in simple landscapes. *J. Phys. A* **30** (1997) 4481.

[Gan86] F. R. Gantmacher. *Matrizentheorie.* Springer, Berlin, 1986.

[Gar] T. Garske and U. Grimm. A maximum principle for the mutation–selection equilibrium of nucleotide sequences. In preparation.

[Gra01] R. M. Gray. *Toeplitz and Circulant Matrices: A review.* 2001. `http://ee.stanford.edu/~gray/toeplitz.pdf`

[Her01] J. Hermisson, H. Wagner, and M. Baake. Four-state quantum chain as a model of sequence evolution. *J. Stat. Phys.* **102** (2001) 315–343.

[Her02] J. Hermisson, O. Redner, H. Wagner, and E. Baake. Mutation–selection balance: Ancestry, load, and maximum principle. *Theor. Pop. Biol.* **62** (2002) 9–46.

[Heu92] H. Heuser. *Funktionalanalysis.* Teubner, Stuttgart, 1992.

[Hew69] E. Hewitt and K. Stromberg. *Real and Abstract Analysis.* Springer, Berlin, 1969.

[Hil82] W. G. Hill. Predictions of response to artificial selection from new mutations. *Genet. Res.* **40** (1982) 255–278.

[Hofa] J. Hofbauer. In preparation.

[Hofb] J. Hofbauer. Private communication.

[Hof85] J. Hofbauer. The selection mutation equation. *J. Math. Biol.* **23** (1985) 41–53.

[Hof88] J. Hofbauer and K. Sigmund. *The Theory of Evolution and Dynamical Systems.* Cambridge University Press, Cambridge, 1988.

[Jag75] P. Jagers. *Branching Processes with Biological Applications.* Wiley, Bath, 1975.

[Jör70] K. Jörgens. *Lineare Integraloperatoren.* Teubner, Stuttgart, 1970.

[Kal97] O. Kallenberg. *Foundations of Modern Probability.* Springer, New York, 1997.

[Kar75] K. S. Karlin and H. M. Taylor. *A first course in stochastic processes.* 2nd ed. Academic Press, San Diego, 1975.

[Kar81] K. S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes.* Academic Press, San Diego, 1981.

[Kat80] T. Kato. *Perturbation Theory for Linear Operators.* Springer, Berlin, 1980.

[Kau87] S. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128** (1987) 11–45.

[Kau93] S. A. Kauffman. *The Origin of Order.* Oxford University Press, New York, 1993.

[Kes66] H. Kesten and B. P. Stigum. A limit theorem for multidimensional Galton–

Watson processes. *Ann. Math. Statist.* **37** (1966) 1211–1233.

[Kim65] M. Kimura. A stochastic model concerning the maintenance of genetic variability in quantitative characters. *Proc. Natl. Acad. Sci. U.S.A.* **54** (1965) 731–736.

[Kim66] M. Kimura and T. Maruyama. The mutational load with epistatic gene interactions in fitness. *Genetics* **54** (1966) 1337–1351.

[Kon88] A. S. Kondrashov. Deleterious mutations and the evolution of sexual reproduction. *Nature* **336** (1988) 435–440.

[Kra72] M. A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Ya. B. Rutitskii, and V. Ya. Stetsenko. *Approximate Solution of Operator Equations.* Wolters-Noordhoff, Groningen, 1972.

[Kre99] R. Kress. *Linear Integral Equations.* 2nd ed. Springer, Heidelberg, 1999.

[Kur97] T. Kurtz, R. Lyons, R. Pemantle, and Y. Peres. A conceptual proof of the Kesten–Stigum theorem for multi-type branching processes. In: K. B. Athreya and P. Jagers (eds.), *Classical and Modern Branching Processes*, (pp. 181–185). Springer, New York, 1997.

[Lan93] S. Lang. *Real and Functional Analysis.* 3rd ed. Springer, New York, 1993.

[Lan99] S. Lang. *Complex Analysis.* 4th ed. Springer, New York, 1999.

[Leu86] I. Leuthäusser. An exact correspondence between Eigen's evolution model and a two-dimensional Ising system. *J. Chem. Phys.* **84** (1986) 1884–1885.

[Leu87] I. Leuthäusser. Statistical mechanics on Eigen's evolution model. *J. Stat. Phys.* **48** (1987) 343–360.

[Lin77] J. Lindenstrauss and L. Tzafriri. *Classical Banach Spaces I. Sequence Spaces.* Springer, Berlin, 1977.

[Lin79] J. Lindenstrauss and L. Tzafriri. *Classical Banach Spaces II. Function Spaces.* Springer, Berlin, 1979.

[Mar76] R. H. Martin. *Nonlinear Operators and Differential Equations in Banach Spaces.* Wiley, New York, 1976.

[Mor95] K. E. Morrison. Spectral approximation of multiplication operators. *New York J. Math.* **1** (1995) 75–96.

[O'B85] P. O'Brien. A genetic model with mutation and selection. *Math. Biosci.* **73** (1985) 239–251.

[Oht73] T. Ohta and M. Kimura. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* **20** (1973) 201–204.

[Oht83] T. Ohta. On the evolution of multigene families. *Theor. Pop. Biol.* **23** (1983) 216–240.

[Ped89] G. K. Pedersen. *Analysis Now.* Revised ed. Springer, New York, 1989.

[Ree80] M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis.* Academic Press, San Diego, 1980.

[Rem98] R. Remmert. *Classical Topics in Complex Function Theory.* Springer, New York, 1998.

[Rud86] W. Rudin. *Real and Complex Analysis.* 3rd ed. McGraw-Hill, New York, 1986.

[Rud91] W. Rudin. *Functional Analysis.* 2nd ed. McGraw-Hill, New York, 1991.

[Sch74] H. H. Schaefer. *Banach Lattices and Positive Operators.* Springer, Berlin, 1974.

[Shi96] A. N. Shiryaev. *Probability.* 2nd ed. Springer, New York, 1996.

[Shp02] M. Shpak and K. Atteson. A survey of unequal crossover systems and their mathematical properties. *Bull. Math. Biol.* **64** (2002) 703–746.

[Swe82] J. Swetina and P. Schuster. Self–replication with errors. A model for polynucleotide replication. *Biophys. Chem.* **16** (1982) 329–345.

[Tak81] N. Takahata. A mathematical study on the distribution of the number of repeated genes per chromosome. *Genet. Res.* **38** (1981) 97–102.

[Tar92] P. Tarazona. Error threshold for molecular quasispecies as phase transition: From simple landscapes to spin glass models. *Phys. Rev.* **A45** (1992) 6038–6050.

[Tho74] C. J. Thompson and J. L. McBride. On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules. *Math. Biosci.* **21** (1974) 127–142.

[Wal87] J. B. Walsh. Persistence of tandem arrays: Implications for satellite and simple-sequence DNA's. *Genetics* **115** (1987) 553–567.

[Wal98] W. Walter. *Ordinary Differential Equations.* Springer, New York, 1998.

[Wer00] D. Werner. *Funktionalanalysis.* 3rd ed. Springer, Berlin, 2000.

[Wie97] T. Wiehe. Model dependency of error thresholds: The role of fitness functions and contrasts between the finite and infinite sites models. *Genet. Res. Camb.* **69** (1997) 127–136.

[Wil65] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Oxford University Press, Oxford, 1965. Reprinted 1992.

[Yos80] K. Yosida. *Functional Analysis.* 6th ed. Springer, Berlin, 1980.

Notation index

Chapter I

Chapter II

Chapter III

Subject index

Acknowledgments

First I want to express my gratitude to Michael Baake for being my thesis advisor, but also for his cooperation on the unequal crossover models and—last but not least—his flexibility regarding my location. Special thanks also go to my co-advisor Ellen Baake, who helped me a great deal in understanding more of biology and mathematics. I furthermore want to cordially thank Joachim Hermisson for asking me to participate in the project on the discrete mutation–selection models, for numerous stimulating discussions, his hospitality during my visit to Yale University in December 2000, and helpful comments on the manuscript of this thesis. I am glad to have had such nice colleagues in my workgroup as Ulrich Hermisson, Moritz Höffe, Christoph Richard, Bernd Sing, and the above-mentioned.

I thank Tini Garske and Wolfgang Angerer for several discussions and helpful comments on the manuscript. To Reinhard Bürger I am grateful for his careful reading of our article on the discrete mutation–selection models [Her02] and for raising the question about similar results for continuum-of-alleles models, which ultimately led to Chapter II. I am indebted to Manfred Wolff for providing working space and library access at Tübingen University and for a number of discussions.

Further thanks go to the Erwin Schrödinger International Institute for Mathematical Physics in Vienna for support during a stay in December 2002. A PhD scholarship by the Studienstiftung des deutschen Volkes is gratefully acknowledged.

Lebenslauf

Name:	Oliver Redner
Geburtsdatum/-ort:	19.11.1972, Hannover
Eltern:	Klaus Dieter Redner und Ursula Redner, geb. Bobe

1979–1983	Grundschule Tegelweg, Hannover
1983–1985	Orientierungsstufe Lüerstraße, Hannover
1985–1992	Kaiser-Wilhelm-Gymnasium, Hannover, Abitur (Note 1,1)
1989–1996	freier Programmierer bei der ConSoft GmbH, Hannover
7/1992–9/1993	Zivildienst in Hannover
10/1993–7/1999	Physik-Studium, Universität Tübingen
26.3.1996	Diplom-Vorprüfung (Note "sehr gut")
6/1996–7/1999	Stipendiat der Studienstiftung des deutschen Volkes
9/1996–5/1997	University of Massachusetts, Amherst, USA (Auslandsaufenthalt)
10/1998	Beginn der Diplomarbeit "Effiziente Simulation periodischer und nichtperiodischer Ising-Modelle am kritischen Punkt" bei PD Dr. Michael Baake am Institut für Theoretische Physik, Universität Tübingen
10/1998–7/1999	wissenschaftliche Hilfskraft am Mathematischen Institut, Universität Tübingen, bei Dr. Helmut Fischer
1.7.1999	Diplom (Note "mit Auszeichnung")
9/1999–9/2000	wissenschaftlicher Mitarbeiter im DFG-Schwerpunkt "Quasikristalle" am Institut für Theoretische Physik, Universität Tübingen, bei M. Baake
10/2000–3/2001	Beginn der Promotion am Institut für Theoretische Physik, Universität Tübingen
10/2000–3/2003	Stipendiat der Studienstiftung des deutschen Volkes
10/2000–2/2001	wissenschaftliche Hilfskraft am Mathematischen Institut, Universität Tübingen, bei M. Baake
4/2001–3/2003	Fortführung der Promotion am Institut für Mathematik und Informatik, Universität Greifswald
6/2001–9/2002	wissenschaftliche Hilfskraft am Institut für Mathematik und Informatik, Universität Greifswald

Fremdsprachen:	Latein (Kl. 5–13, Großes Latinum)
	Englisch (Kl. 7–11, Auslandsaufenthalte, Veröffentlichungen)
	Altgriechisch (Kl. 9–10)
	Französisch (Kl. 11–13)
	Intensiv-Kurs Italienisch (3/1999)
Weitere Interessen:	Klavierspielen und -unterrichten