WKB theory of large deviations in stochastic populations

Topical Review

WKB theory of large deviations in stochastic populations

Michael Assaf and Baruch Meerson

Racah Institute of Physics, Hebrew University of Jerusalem, Jerusalem 91904, Israel

E-mail: michael.assaf@mail.huji.ac.il and meerson@mail.huji.ac.il

Abstract

Stochasticity can play an important role in the dynamics of biologically relevant populations. These span a broad range of scales: from intra-cellular populations of molecules to population of cells and then to groups of plants, animals and people. Large deviations in stochastic population dynamics—such as those determining population extinction, fixation or switching between different states—are presently in a focus of attention of statistical physicists. We review recent progress in applying different variants of dissipative WKB approximation (after Wentzel, Kramers and Brillouin) to this class of problems. The WKB approximation allows one to evaluate the mean time and/or probability of population extinction, fixation and switches resulting from either intrinsic (demographic) noise, or a combination of the demographic noise and environmental variations, deterministic or random. We mostly cover well-mixed populations, single and multiple, but also briefly consider populations on heterogeneous networks and spatial populations. The spatial setting also allows one to study large fluctuations of the speed of biological invasions. Finally, we briefly discuss possible directions of future work.

Keywords: stochastic population dynamics, stochastic epidemic models, large deviations, population extinction and fixation, genetic switches, biological invasions, master equation, WKB methods

(Some figures may appear in colour only in the online journal)

1. Introduction

It was realized long ago that stochasticity can play an important role in the dynamics of biologically relevant populations. These span a broad range of scales: from intra-cellular populations of molecules to population of cells and then to groups of plants, animals and people [1–12]. The stochasticity, or noise may have different origins. Two main types of noise have

been identified: intrinsic, or demographic noise and extrinsic, or environmental noise. The demographic noise reflects natural 'quantization' of individuals the population is made of, and random character of elemental transitions such as births, deaths, interactions and movement of the individuals. Environmental variations (not necessarily stochastic) are usually described in terms of time-dependent elemental transition rates. If one ignores the noise, a steady-state population size corresponds to an attractor in the space of population sizes as described by a *deterministic* rate equation, or equations. Taking the stochasticity into account, one observes random excursions of the population sizes around the attractor. For *established* populations the typical excursions around the attractor are small. However, from time to time a rare large excursion occurs, which may lead to extinction of one of more populations, or to a population switch to the vicinity of another attractor. As a result, stochasticity can make deterministically stable attractors *metastable*. In such situations it is interesting, and often biologically important, to determine the mean time to extinction or switch, when starting from the vicinity of an attractor. This review will mostly deal with problems of this type. We will confine ourselves to individual-based Markov population models and not consider continuum models of population dynamics, based on Langevin-type equations, see e.g. [13]. The fact that typical fluctuations are small implies the presence of a small parameter, usually coming from a disparity of the elemental transition rates. The WKB approximations—the focus of this review—utilize this small parameter in a smart way in order to approximately solve the master equation which is considered to be the exact model of the population dynamics. There are three recent reviews on closely related subjects. The review [14] on the WKB approximation in stochastic population models was primarily intended for population biologists and ecologists, and it dealt only with extinction of well-mixed populations. The review [15] provided a comprehensive and pedagogic introduction to the path integral representation of master equations and to approximate methods of solution. Finally, the review [16] is a survey of different mathematical methods of analysis of biological switching processes in a variety of systems, at both the genotypic and phenotypic levels.

In a well-mixed single population—the simplest paradigm of stochastic population dynamics—the only generic type of attractor of the deterministic rate equation is a stable fixed point. As we will see, even in this simple situation there are two different scenarios of population extinction caused by demographic noise. Scenario A is observed in the absence of an Allee effect, when the population is monostable. Scenario B corresponds to bistability caused by an Allee effect. In the context of population switches due to a weak noise one always deals with bistability.

When there are two interacting populations (predator and prey, competition or symbiosis, susceptible and infected populations, etc), a generic attractor can be either a stable fixed point (a node or a focus), or a stable limit cycle, and we will review these cases separately. As the number of interacting populations increases, additional types of attractors may appear, including chaotic attractors [17]. Noise-induced escape from chaotic attractors has been studied theoretically [18, 19] and experimentally [20] for continuous noisy systems, as described by Langevin equations, but not for stochastic populations.

The WKB (Wentzel–Kramers–Brillouin) approximation is best known to physicists in the context of quantum mechanics. In the time-dependent WKB theory a WKB ansatz leads to an approximate description of the time-dependent wave function as described by the non-stationary Schrödinger equation: for example, for a (quasi-classical) quantum particle in a time-dependent potential, see [21], chapter III, section 17. In its turn, the stationary WKB theory can be used for approximate calculations of the eigenvalues and eigenfunctions corresponding to highly excited states as described by the stationary Schrödinger equation, for evaluating the

(small) probability of tunneling through a potential barrier, and for other purposes, see chapter VII of [21]. Similarly, there are time-dependent and stationary versions of the WKB approximation in stochastic problems as we will see below.

For stochastic classical systems with a *continuous* space of states a WKB theory was developed by Freidlin and Wentzel [22], Dykman [23] and Graham [24]. As populations are naturally 'quantized'—they have an integer number of individuals—the WKB approximation here has some important differences. As of present, there are two main strategies of applying the WKB approximation to stochastic populations. In the more straightforward 'real-space' WKB method (where by 'real space' is actually meant the space of population sizes) one applies a WKB ansatz directly to the master equation, or to the equation describing the stationary or quasi-stationary distribution [25–57]. In the 'momentum space' WKB method one first derives an exact evolution equation—a linear partial differential equation (PDE)—for the probability generating function. For example, for a well-mixed single population the probability generating function is defined as [58, 59]

$$G(p,t) = \sum_{n=0}^{\infty} p^n P_n(t), \tag{1}$$

where $P_n(t)$ is the probability of observing n individuals at time t and p is an auxiliary variable. Then one applies a WKB ansatz to the evolution equation for $G(p,t)$ or to the stationary equation describing the eigenstates of $G(p,t)$ [60–70]. For established populations a small parameter required for the WKB approximation is $1/N$, where $N \gg 1$ is a typical population size characterizing the (quasi)stationary distribution. In well-mixed single populations each of the two WKB methods, in combination with additional perturbation techniques, using the same small parameter $1/N$, yields accurate and controllable results, in the leading and subleading orders of N, for the mean time to extinction/switch. Even the leading-order WKB results, which miss pre-exponential factors, are usually much more accurate than the results obtained with the more traditional 'diffusion approximation' based on the van Kampen system-size expansion [5]. Indeed, as was shown in many studies [30, 60, 62, 63, 71, 72], the mean time to extinction/switch, obtained with the diffusion approximation, usually involves an error that is exponentially large in N.

For multiple populations, even the leading-order results for the mean time to extinction/switch are, in general, unavailable in analytical form. Still, the WKB method proves very useful here, as it yields, albeit in a numerical form, *the optimal path*: a special trajectory of the system in the phase space of population sizes and the conjugate momenta that gives a dominant contribution to the specified large deviation such as extinction or switch. In its turn, the mean time to extinction/switch is given (again, in a numerical form) by the 'classical action' evaluated along the optimal path. In this class of problems analytical progress is possible if, besides N, there is an additional small parameter in the system, coming from an additional disparity in the elemental transition rates [31, 34, 38, 47, 65, 73].

As we already mentioned, environmental/extrinsic variations are usually taken into account by allowing some of the elemental transition rates to vary with time. In some cases these variations are deterministic: describing seasonality of births and deaths, 'catastrophes' and other effects in population biology, and vaccination in epidemiology. In other cases the variations can be viewed as stochastic, reflecting multiple concurrent mechanisms. In the leading order of the WKB approximation these problems are similar to multi-population problems. Here too the WKB theory yields the optimal path of the population to the specified large deviation [50, 64, 66, 67, 74–76] and, for a stochastic rate variation, the optimal realization of the environmental noise that gives a dominant contribution to this large deviation [50, 74, 75].

Figure 1. Scenario A. The full circle denotes an attracting fixed point of the deterministic rate equation, the full square denotes an absorbing state of the stochastic process. Left panel: extinction. Right panel: fixation.

Figure 2. Scenario B. The full circle denotes an attracting fixed point, the full square denotes an absorbing state, and the empty circle denotes a repelling fixed point. Left panel: extinction. Right panel: switching.

When the number of interacting sub-populations increases, their topological heterogeneity becomes very important in determining the large deviation properties. This is a largely unexplored subject, and we will review some recent work in this direction [77].

Introducing migration of populations in space (traditionally modeled as random walk of the individuals) makes the large deviation problems both richer and more difficult. Here the WKB method, once applicable, provides a convenient classical-field-theory framework for studying large deviations: from population extinction [60, 78] to large velocity fluctuations of biological invasion fronts [79–81].

In the last few years the WKB approximation in stochastic population dynamics has been adopted by a growing number of practitioners. This review (which may have a natural bias toward our own work) does not attempt to cover all of them.

2. Extinction, fixation and switching in established single populations

In this section we will briefly review the two WKB methods—the real-space and the momentum-space—on several examples of escape from a long-lived metastable state corresponding to an established well-mixed single population. The elemental transitions do not need to be single-step: they can involve transitions between a state with n individuals and a state with $n + r$ individuals, where $r = \pm 1, \pm 2, \ldots$. These transitions occur with the rate $W_r(n)$, so the master equation reads

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = \sum_r \left[W_r(n - r) P_{n-r}(t) - W_r(n) P_n(t) \right], \tag{2}$$

where $P_{k<0} = 0$.

Typically, there are two escape routes, or escape scenarios, from a metastable state of a single population [35]. Let us denote the attracting fixed point, in the vicinity of which the population resides, by n_*. In escape scenario A there exist one or two adjacent repelling fixed points of the deterministic rate equation that are also absorbing states of the stochastic dynamics. A single repelling point $n = n_u = 0$ corresponds to extinction; two repelling points correspond to fixation, see figure 1.

In escape scenario B, in addition to the repelling point $n_u > 0$, which is now *non*-absorbing, there exists a second attracting fixed point of the deterministic rate equation, $n = n_s$, see figure 2. If $n_s = 0$ and absorbing, this is extinction. If $n_s > 0$ and non-absorbing, this is switching from n_* to n_s. Each of the scenarios A and B can be dealt with by the real-space WKB approach, while the momentum-space approach is limited to scenario A [68].

Let us derive a general approximate expression for the mean time to escape. After a short relaxation time t_r, determined by the deterministic rate equation, the population gets established as a long-lived metastable distribution peaked near $n = n_*$. The metastable distribution very slowly decays in time because of the probability 'leakage' into the absorbing state(s). The long-time decay is characterized by the lowest positive eigenvalue $1/\tau$ of the master equation (2), so that $\tau \gg t_r$. That is, at times $t \gg t_r$, one has [27, 30, 32, 33, 35]

$$P_{n>0}(t \gg t_r) \simeq \pi_n \mathrm{e}^{-t/\tau}, \quad P_0(t \gg t_r) \simeq 1 - \mathrm{e}^{-t/\tau}. \tag{3}$$

Here we have assumed for concreteness a single absorbing state at $n = 0$, corresponding to extinction. The case of fixation will be dealt with shortly. The function π_n ($n = 1, 2, \ldots$) describes the quasi-stationary distribution (QSD) of the population. For metastable populations the decay time τ is an accurate approximation to the mean time to extinction (MTE) or fixation. Using equation (3), one arrives at an eigenvalue problem for the QSD π_n, $n = 1, 2 \ldots$:

$$\sum_r \left[W_r(n - r)\pi_{n-r} - W_r(n)\pi_n \right] = -\pi_n/\tau. \tag{4}$$

At large N the eigenvalue $1/\tau$ is exponentially small, see below. Therefore, the term in the right hand side of equation (4) can be neglected, and we arrive at a quasi-stationary equation

$$\sum_r \left[W_r(n - r)\pi_{n-r} - W_r(n)\pi_n \right] = 0, \quad n = 1, 2, \ldots. \tag{5}$$

Once the QSD is known, we can plug equation (3) into equation (2) for $n = 0$, and obtain the MTE from the relation

$$1/\tau = \sum_{r>0} W_r(r)\pi_r. \tag{6}$$

Equation (6) holds for both scenarios A and B. The QSD itself, however, is scenario-dependent, and this ultimately determines the MTE τ. In the following we will derive the expressions for the QSD for each of the two escape scenarios.

2.1. Extinction and fixation: scenario A

In section 2.1.1 we will determine the QSD in the case of extinction via scenario A. Section 2.1.2 will deal with fixation.

2.1.1. Extinction. We will illustrate some general results, that we will present shortly, on a typical example: a variant of the stochastic Verhulst model [82] with the (rescaled) birth and death rates

$$W_{+1}(n) = Bn, \quad W_{-1}(n) = n + B\frac{n^2}{N}, \quad N \gg 1. \tag{7}$$

The quadratic corrections account for competition for resources. For this model, the master equation (2) is

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = W_{+1}(n - 1)P_{n-1}(t) + W_{-1}(n + 1)P_{n+1}(t) - [W_{+1}(n) + W_{-1}(n)]P_n(t),$$
$$\frac{\mathrm{d}P_0(t)}{\mathrm{d}t} = W_{-1}(1)P_1(t). \tag{8}$$

The deterministic rate equation can be obtained by multiplying both sides of equation (8) by n and summing over all n's. Using the mean-field assumption $\langle n^2 \rangle = \langle n \rangle^2$, this procedure yields [58, 59]

$$\dot{\bar{n}} = (B-1)\bar{n} - B\frac{\bar{n}^2}{N}, \tag{9}$$

where $\bar{n}$ denotes the mean population size. This equation (that also emerges from the WKB formalism) has two fixed points: at $\bar{n} = n_u = 0$ and at $\bar{n} = n_* = N(B-1)/B$. We will assume that $B > 1$, so these fixed points are repelling and attracting, respectively. As a result, for $B > 1$ the population size will flow to the attracting fixed point at n_* when starting from any nonzero value of n at $t = 0$. The metastable population distribution is peaked at about $n = n_*$, whereas the repelling fixed point $n = 0$ is an absorbing state of the stochastic process.

Let us determine the QSD by using the real-space WKB method. To this end we express the QSD in terms of the rescaled population size $q = n/N$ and look for the solution of equation (5) at $n \gg 1$ by making the WKB ansatz [27, 30, 32, 33, 35]

$$\pi_n \equiv \pi(q) \simeq A e^{-NS(q)-S_1(q)-\cdots}, \tag{10}$$

where $S(q)$ and $S_1(q)$ are assumed to be $O(1)$, and a constant prefactor A is introduced for convenience. The leading-order calculations, where one neglects S_1, are straightforward as we show shortly. The sub-leading order calculations demand some effort. This is because the subleading WKB solution for the QSD has to be complemented by a recursive solution of the master equation in the vicinity of the absorbing state. In the case of extinction the subleading WKB solution has to be matched with a recursive solution valid at $n = O(1)$ [30, 35], see below.

We plug the WKB ansatz (10) into equation (5) and, assuming $n \gg 1$, Taylor-expand the functions of $q \pm 1/N$ around q. In the leading and subleading orders we obtain

$$S(q) = \int^q p_a(q) \mathrm{d}q, \quad S_1(q) = \frac{1}{2} \ln[w_+(q)w_-(q)]. \tag{11}$$

Here $w_+(q) = W_+(n)/N = Bq$ and $w_-(q) = W_-(n)/N = q(1 + Bq)$ are the rescaled transition rates, see equation (7). Furthermore, $p_a(q) = \ln[w_-(q)/w_+(q)] = \ln[(1 + Bq)/B]$ is the so called fast-mode WKB solution [35], see figure 3. This solution describes the optimal path: the nontrivial zero-energy trajectory of the effective classical Hamiltonian

$$H(q,p) = w_+(q)\left(e^p - 1\right) + w_-(q)\left(e^{-p} - 1\right), \tag{12}$$

where $p = \mathrm{d}S/\mathrm{d}q$ is the momentum. The constant A in equation (10) can be found by normalizing the QSD in the vicinity of the attracting fixed point $q_* = n_*/N = 1 - 1/B$, where the QSD can be approximated by a Gaussian. This yields

$$A = \sqrt{\frac{S''(q_*)}{2\pi N}} e^{NS(q_*)+S_1(q_*)}. \tag{13}$$

(Note that there is an additional, slow-mode WKB solution [35] for which $S(q) = 0$ and $S_1(q) = \ln[w_+(q) - w_-(q)]$. In scenario A this solution gives a negligible contribution in the entire WKB region, so it should be discarded. In scenario B, however, this solution plays an important role, see below.) Plugging equations (11) and (13) into equation (10), we obtain the WKB approximation for the QSD for the Verhulst model [35]:

$$\pi_n = \pi(q) = \frac{B-1}{\sqrt{2\pi NBq^2(1+Bq)}} e^{-NS(q)}, \tag{14}$$
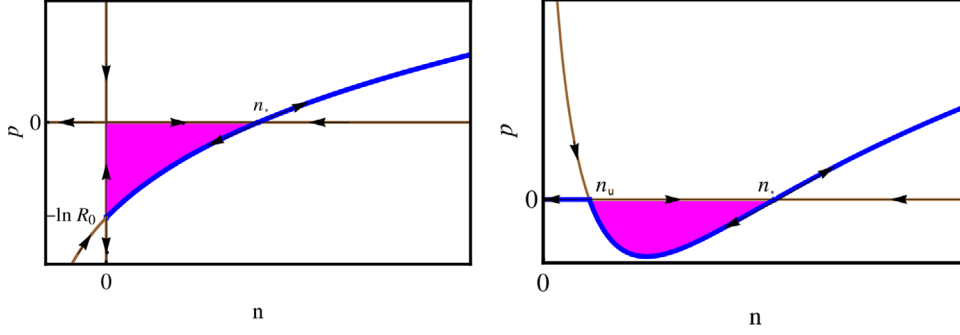
Figure 3. Extinction scenarios A and B. Left panel: scenario A (no Allee effect): the repelling fixed point $\bar{n} = 0$ is an absorbing state of the stochastic process. The stochastic extinction occurs via a large fluctuation which brings the established population from a vicinity of the attracting fixed point $\bar{n} = n_*$ directly to the absorbing state $n = 0$. Right panel: scenario B (Allee effect): the absorbing state at $n = 0$ is an attracting fixed point. The stochastic extinction occurs via a large fluctuation which brings the established population from a vicinity of the *next* attracting fixed point $\bar{n} = n_*$ to a vicinity of the repelling fixed point $\bar{n} = n_u$. From there the population size flows 'downhill' to the absorbing state $n = 0$ almost deterministically. The thick curves show the corresponding optimal paths $p = p(q)$. Reprinted figure with permission from [35], copyright 2015 by the American Physical Society.

where

$$S(q) = 1 - \frac{1}{B} - q + \left(q + \frac{1}{B} \right) \ln \left(q + \frac{1}{B} \right). \tag{15}$$

The action function $S(q)$ is depicted in figure 4.

The approximation (14) and (15) is invalid at $n = O(1)$, or $q = O(1/N)$. Here one has to solve the quasi-stationary master equation recursively. Fortunately, one can neglect the non-linear terms in the transition rates [30, 35]. The recursive solution is

$$\pi_n = \frac{(B^n - 1)}{\tau (B - 1) n}, \tag{16}$$

where τ is the MTE we are after. Matching this recursive solution with the WKB solution (14) in their joint region of validity $1 \ll n \ll \sqrt{N}$ and using equations (6) and (15), we find [35]

$$\tau = \sqrt{\frac{2\pi B}{N}} \frac{1}{(B-1)^2} \exp \left[\frac{N}{B} \left(B - 1 - \ln B \right) \right] \equiv \sqrt{\frac{2\pi B}{N}} \frac{1}{(B-1)^2} e^{NS(0)}. \tag{17}$$

As expected, τ is exponentially large in N. As one can clearly see now, the leading-order WKB action $S(q)$ plays the role of (and sometimes called) the non-equilibrium potential. It describes an effective exponential barrier to extinction. The pre-exponential factor is important in this case, as it includes the large parameter $N^{-1/2}$. The presence of a large parameter in the pre-exponent is a typical feature of the extinction scenario A. Figure 5 shows a comparison of the mean extinction rate $1/\tau$ from equation (17) with the mean extinction rate measured in Monte-Carlo simulations at different $N \gg 1$. Very good agreement is observed.

Close to the transcritical bifurcation of the deterministic rate equation, $\delta \equiv B - 1 \ll 1$, but still at $N\delta^2 \gg 1$, equation (17) simplifies to

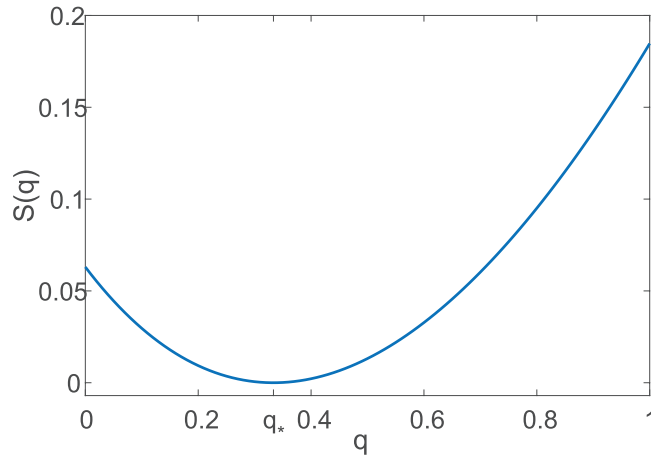$$\tau \simeq \sqrt{\frac{2\pi}{N}} \frac{e^{N\delta^2/2}}{\delta^2}. \tag{18}$$

Figure 4. The action function $S(q)$ for the Verhulst model, given by equation (15), for $B = 1.5$. Here $q_* = 1 - 1/B = 1/3$.
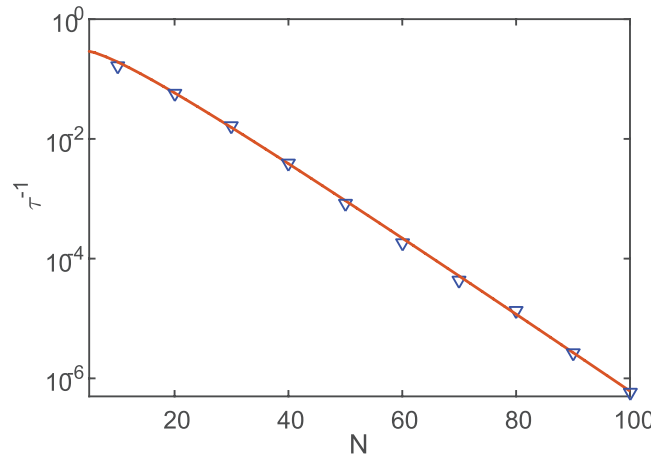

Figure 5. The mean extinction rate τ^{-1} versus N on a semi-log scale for $B = 2$. Line: theoretical result (17), symbols: results of Monte-Carlo simulations.

This close-to-the-bifurcation result is universal: it applies to a whole class of population models that do not include an Allee effect [35].

Now, continuing using the Verhulst model as an example, we will show that the often used van-Kampen system size expansion (also called the diffusion approximation) generally yields an error in the MTE that is exponentially large in N [30, 60, 62, 63, 71, 72]. The van-Kampen system size expansion approximates the exact master equation (8) by a Fokker–Planck equation. The Fokker–Planck equation can be obtained from equation (8) in the following way. One defines $j_\pm(n) = W_{\pm 1}(n)P_n(t)$, and Taylor expands $j_\pm(n \mp 1)$ around n up to the second order. Then, introducing the rescaled population size $q = n/N$, one arrives at the Fokker–Planck equation

$$\frac{\partial P(q,t)}{\partial t} = -\frac{\partial}{\partial q}\left[v(q)P(q,t)\right] + \frac{1}{N}\frac{\partial^2}{\partial q^2}\left[d(q)P(q,t)\right]. \tag{19}$$

Here

$$v(q) = w_+(q) - w_-(q) = q(B - 1 - Bq)$$

is the deterministic drift velocity, and

$$d(q) = w_+(q) + w_-(q) = q(B + 1 + Bq)$$

is the diffusivity that describes an effective multiplicative noise. Now we continue exploiting the large parameter $N \gg 1$. We make the ansatz $P(q, t) = \pi(q)e^{-t/\tau}$ and use the WKB approximation, see equation (10), to analyze the quasi-stationary equation for $\pi(q)$. This procedure yields, in the leading order in N, a Hamilton–Jacobi equation $H(q, dS/dq) = 0$ with the Hamiltonian

$$H(q, p) = p \left[v(q) + \frac{d(q)\, p}{2} \right]. \tag{20}$$

The resulting action function is

$$S^{\mathrm{FP}}(q) = -\int_{q_*}^{q} \frac{2v(q')}{d(q')}\, dq' = 2 \left[q - 1 + \frac{1}{B} - 2\ln \frac{(1 + B + Bq)}{2B} \right]. \tag{21}$$

This expression (which is then multiplied by a large N inside the exponent of the MTE) differs from the asymptotically exact expression (15) thus invalidating the Fokker–Planck equation as a controlled approximation for the purpose of dealing with large deviations.

The van-Kampen system size expansion may become accurate close to bifurcation points of the deterministic model. In our Verhulst example, the bifurcation is at $B = 1$. At $B - 1 \ll 1$, the stable fixed point is $q_* \simeq B - 1 \ll 1$. Therefore we can expand $v(q)$ and $d(q)$ in the vicinity of $q = 0$, which yields

$$H \simeq pq \left[v'(0) + \frac{v''(0)}{2}q + \frac{d'(0)}{2}p \right] \simeq pq(B - 1 - q + p). \tag{22}$$

As one can check, this Hamiltonian coincides with the asymptotically exact WKB Hamiltonian (12) close to the bifurcation at $B = 1$. Naturally, the zero-energy activation trajectory,

$$p(q) = -\frac{2v'(0) + qv''(0)}{d'(0)} = q + 1 - B, \tag{23}$$

is also the same as the one following from the Hamiltonian (12), and $S^{\mathrm{FP}}(0)$ is given by $(1/2)(B - 1)^2$, which coincides with the exponential term in equation (18). Performing the subleading-order calculations, one arrives at a pre-exponent which also coincides with that of equation (18) (see [35] for details).

We now present, using the same example of the Verhulst model, the momentum-space WKB method. Multiplying both sides of the master equation (8) by p^n and summing over all p one arrives at an exact evolution equation for the probability generating function $G(p, t)$ from equation (1):

$$\frac{\partial G}{\partial t} = (p - 1) \left\{ \left[Bp - \left(1 + \frac{B}{N} \right) \right] \frac{\partial G}{\partial p} - \frac{Bp}{N} \frac{\partial^2 G}{\partial p^2} \right\}. \tag{24}$$

One way of using the WKB approximation would be to look for the time-dependent solution as $G(p, t) = \exp[-S(p, t)]$ and neglecting, in the leading WKB order, the second derivative $\partial_p^2 S$ [60]. This would suffice if we limit ourselves to the leading WKB order. For more accurate calculations, that yield pre-exponents, it is convenient to first exploit the known

spectral properties of the metastability problem [61–63, 68]. Indeed, the metastable solution, corresponding to the established population whose distribution is sharply peaked around $n_* = N(1 - 1/B)$, is described by the lowest excited eigenmode $\phi(p)$ of the differential operator in the right hand side of equation (24): the mode that decays with the rate $1/\tau$,

$$G(p, t) = 1 - \phi(p)e^{-t/\tau}. \tag{25}$$

At $t \to \infty$ one obtains $G(p, t) = 1$, corresponding to population extinction, $P_n(t \to \infty) = \delta_{n,0}$. We plug the ansatz (25) into equation (24) and arrive at an ordinary differential equation (ODE) for $\phi(p)$:

$$(p - 1) \left\{ \left[Bp - \left(1 + \frac{B}{N} \right) \right] \phi'(p) - \frac{Bp}{N} \phi''(p) \right\} = \frac{\phi}{\tau}, \tag{26}$$

where the prime denotes the derivative with respect to the argument. The probability conservation yields $G(1, t) = 1$ (see equation (1)), therefore $\phi(1) = 0$. Analyticity of $\phi(p)$ in the vicinity of $p = 0$ brings about an additional condition:

$$\left(1 + \frac{B}{N} \right) \phi'(0) - \frac{\phi(0)}{\tau} = 0$$

which, in view of the expected exponential smallness of $1/\tau$, can be replaced by $\phi'(0) = 0$. The 'self-generated' boundary conditions $\phi'(0) = \phi(1) = 0$ define the natural interval $0 \leqslant p \leqslant 1$ on which the boundary value problem for equation (26) should be considered [61–63, 68]. In order to find the MTE, we can consider two separate regions and match the asymptotic solutions in their joint region of validity. In the bulk—that is, not too close to $p = 1$—we can assume that $\phi(p)$ is almost constant, and look for a perturbative solution

$$\phi(p) = 1 + \delta\phi(p), \quad \delta\phi(p) \ll 1. \tag{27}$$

Plugging this ansatz into equation (26), and defining $u(p) = \phi'(p)$, the perturbative solution in the bulk can be written as

$$u^b(p) = \frac{N}{B\tau p} e^{-NS(p)} \int_0^p \frac{e^{NS(p)}}{p - 1} dp, \tag{28}$$

where $S(p) = (1/B) \ln p - p + 1$, and the boundary condition $u(0) = 0$ is satisfied.

In the boundary-layer region $1 - p \ll 1$ we notice that $\phi(p)$ rapidly falls from a value close to unity to zero over a narrow region of size $O(N^{-1/2}) \ll 1$. In the boundary layer, the approximate ODE reads

$$\left[Bp - \left(1 + \frac{B}{N} \right) \right] u(p) - \frac{Bp}{N} u'(p) = 0. \tag{29}$$

Here we look for a WKB-type solution $u(p) = a(p)e^{-NS(p)}$ where $S(p)$ is the action in the momentum space, and $a(p)$ is an amplitude. Plugging this ansatz in equation (29), we demand that the terms cancel each other separately in the zeroth and first order in $1/N$. What is the boundary condition for $u(p)$ at $p = 1$? At times $t \ll \tau$ we have, on the one hand, $\partial_p G(p = 1, t) = n_* = N(B - 1)/B$ and, on the other hand, $\partial_p G(p = 1, t) = -u(1)$, so $u(1) = -n_*$. As a result, the boundary-layer solution is

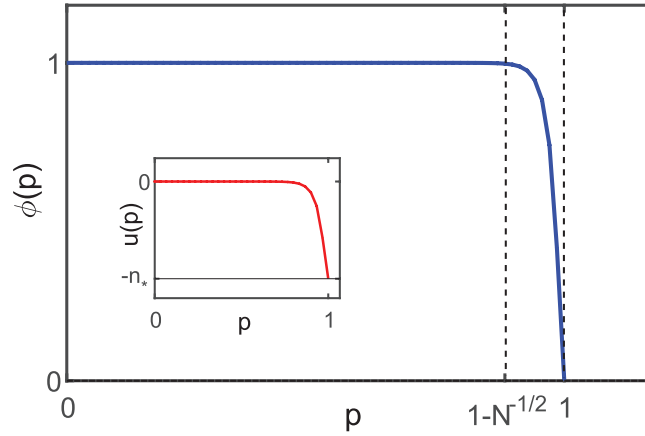$$u^{bl}(p) = -\frac{n_*}{p} e^{-NS(p)}. \tag{30}$$

Figure 6. An illustration of the function $\phi(p)$ in the region of interest $p \in [0, 1]$. The boundary layer is schematically denoted by the region bounded by the vertical dashed lines. The inset shows $u(p) = \phi'(p)$.

The solutions to $\phi(p)$ and $u(p)$ are illustrated in figure 6.

In order to find the MTE τ, one can match the bulk solution (28) and the boundary-layer solution (30) in their joint region of validity $N^{-1/2} \ll 1 - p \ll 1$ [61–63, 68]. For $1 - p \gg N^{-1/2}$, the integral in the bulk solution (28) can be evaluated by a saddle-point approximation, where the saddle point is found at $p = 1/B$. This gives

$$\int_0^p \frac{e^{NS(p)}}{p - 1} dp \simeq -\sqrt{\frac{2\pi B}{N}} \frac{e^{-NS(1/B)}}{B - 1}. \tag{31}$$

Plugging this into equation (28) and matching with equation (30) we again arrive at equation (17).

Notice that, although the boundary-layer solution (30) has a recognizable WKB form, the actual use of the WKB approximation here was unnecessary. The only approximation used was to linearize equation (26), that is to replace $\phi(p)$ by 1 in the right hand side and neglect the exponentially small term τ^{-1}. This is a typical situation for the elemental transitions leading to a second-order ODE for $\phi(p)$, like equation (26). Yet, already for third-order ODEs (such as the one that appears, for example, when one of the transitions is triple annihilation $3A \to \emptyset$), the WKB approximation becomes indispensable [68].

2.1.2. Fixation. Consider a population consisting of n mutants of type A and $N - n$ wild-type individuals of type B, which can be genes, cells or even animals. What is the probability that the mutants take over (fixate) the entire population, causing the extinction of the wild-type individuals? This question can be addressed in the framework of evolutionary game theory (EGT) which allows to describe how successful strategies spread by imitation or reproduction [83].

In its simplest form, the EGT framework includes two reactions $A + B \to A + A$ at a rate f_A, and $A + B \to B + B$, at a rate f_B. Here, an individual chosen proportionally to its fitness (reproduction potential) produces an identical offspring which replaces a randomly chosen individual, and f_A and f_B denote the fitness of types A and B, respectively. If, in addition, the fitness of A and B depends on the current number of mutants, the deterministic rate equation—the replicator dynamics [83]—describing the mean number of mutants, can be written as

$$\dot{n} = n(N - n)[f_A(n) - f_B(n)]. \tag{32}$$

The simplest form of f_A and f_B, which guarantees an intermediate coexistence state, is when $f_A(n) - f_B(n) \sim n_* - n$, corresponding to the linear Moran model [84]. This yields an attracting point at $n = n_*$ and two repelling fixed points: at $n = 0$ and $n = N$.

The stochastic version of this model is described by the master equation

$$\frac{dP_n(t)}{dt} = W_{+1}(n-1)P_{n-1}(t) + W_{-1}(n+1)P_{n+1}(t) - [W_{+1}(n) + W_{-1}(n)]P_n(t), \tag{33}$$

where $W_{+1}(n) = n(N-n)f_A(n)$ and $W_{-1}(n) = n(N-n)f_B(n)$ [36, 37, 46, 83]. The stochastic dynamics of this problem resemble those of the Verhulst model. After a short relaxation time t_r, the system enters a long-lived metastable coexistence state centered about n_*. Here, however, the system can escape to either the absorbing state $n = 0$ corresponding to *extinction* of the mutants, or to $n = N$ corresponding to *fixation* of the mutants. As a result, the dynamics of the probability distribution function at $t \gg t_r$ satisfy the ansatz [36, 37, 40, 46]

$$P_{1 \leqslant n \leqslant N-1} = \pi_n e^{-t/\tau}, \quad P_0(t) = \varphi(1 - e^{-t/\tau}), \quad P_N(t) = (1 - \varphi)(1 - e^{-t/\tau}), \tag{34}$$

where the mean time to fixation (of either the wild type or the mutants), τ, and the probability of mutant extinction, φ, satisfy

$$\varphi = W_{-1}(1)\pi_1\tau, \quad \tau = [W_{-1}\pi_1 + W_{+1}(N-1)\pi_{N-1}]^{-1}. \tag{35}$$

The solution of this problem can be found using the real-space WKB approach, and follows the same lines as the solution presented above for extinction. As before, the WKB solution breaks down close to the boundaries $n = 0$ and $n = N$, where one has to use recursive solutions by linearizing the reaction rates close to $n = 0$, and $n = N$, respectively; see [36, 37, 40].

The scenario we have presented above contains wild-type individuals and those having a single mutation. However, there exist more complicated scenarios of fixation with multiple mutations, when e.g. a beneficial mutation (with a higher fitness than the wild type's) is obtained from a wild-type species by going through a detrimental mutation (with a lower fitness than the wild type's). Such problems, which include in their simplest form a two-locus genotype space, and in which the negative effects of two single mutations are overcompensated by a higher-fitness double mutant, can also be dealt with by using the real-space WKB approach, see e.g. [85].

2.2. Extinction and switching: scenario B

We will now consider escape scenario B, see figure 2. In section 2.2.1 we will study extinction ($n = n_s = 0$ is absorbing), whereas section 2.2.2 will deal with switching between two metastable states ($n = n_s$ is non-absorbing).

2.2.1. Extinction. Following [35], we use the following set of reactions: binary reproduction $2A \xrightarrow{\lambda} 3A$, the reverse transition $3A \xrightarrow{\sigma} 2A$, and linear decay $A \xrightarrow{\mu} \emptyset$. Here

$$W_1 = \frac{\lambda n(n-1)}{2}, \ W_{-1} = \mu n + \frac{\sigma n(n-1)(n-2)}{6}. \tag{36}$$

The deterministic rate equation is [35]

$$\dot{n} = \frac{\lambda}{2}n^2 - \mu n - \frac{\sigma}{6}n^3. \tag{37}$$

The fixed point at $n = 0$ is now attracting. Furthermore, denoting by $\delta^2 = 1 - 8\sigma\mu/(3\lambda^2) > 0$, and $N = 3\lambda/(2\sigma)$, equation (37) has two additional fixed points $n_u = N(1 - \delta)$ and $n_* = N(1 + \delta)$. An established stochastic population resides in the vicinity of the fixed point n_*. This simple model accounts for the Allee effect that refers to a variety of processes that reduce the per-capita growth rate at small population densities. The Allee effect has been long known in ecology and population biology [86]. Very recently, its importance has been also appreciated in mathematical cancer biology [87].

In contrast to scenario A, where the WKB solution is determined solely by the non-trivial zero energy trajectory (the fast mode), the WKB solution in scenario B includes two modes. The fast-mode WKB solution dominates at $n > n_u$ (but not too close to n_u), whereas the slow-mode WKB solution (see below) dominates at $0 < n < n_u$ (again, not too close to $n = n_u$) [32, 33, 35] , see figure 3. Moreover, because the slow-mode solution diverges at $n = n_u$, one has to go beyond the WKB approximation in a boundary layer $|n - n_u| \ll n_u$ where the fast and slow WKB modes are strongly coupled. Fortunately, in this boundary layer the quasi-stationary master equation can be approximated by a quasi-stationary Fokker–Planck equation [32, 33, 35]. As a result, the QSD at $n \gg 1$ involves *three* distinct asymptotes which need to be matched with one another. Finally, at $n = O(1)$, where the WKB approximation breaks down, a recursive solution for the QSD can be found by neglecting the non-linear terms in the master equation [35]. The recursive solution, however, is only needed if one wants to determine the MTE in a straightforward manner, by using equation (6). There is, however, an important shortcut which does not require the small-n recursive solution [33]. This is because the meta-stable distribution develops a *constant* probability current toward small n. It is this current that determines the escape rate from the metastable state $n = n_*$. One can determine this current in a close vicinity of $n = n_u$, without the need to address the region of $n = O(1)$. The resulting calculations yield a general expression for the MTE [35], which turns out to be the same as the mean time to switch [33], and even as the mean time to population explosion if the attracting fixed point is at $n = \infty$ [32]. In our example of the three reactions with an Allee effect the resulting MTE is the following:

$$\tau = \frac{\pi(1 - \delta)}{\mu\delta} e^{N\Delta S}, \quad \Delta S = 2\left(\delta - \sqrt{1 - \delta^2} \arctan \frac{\delta}{\sqrt{1 - \delta^2}}\right). \qquad (38)$$

As in scenario A, τ is exponentially large in N. Now, however, the pre-exponential factor is independent of N. In figure 7 one can see a comparison between the theoretical formula (38) and numerical Monte-Carlo simulations, and excellent agreement is observed.

Close to the saddle-node bifurcation of the deterministic rate equation, $\delta \ll 1$, but still at $N\delta^3 \gg 1$, equation (38) becomes [33, 35]

$$\tau = \frac{\pi}{\mu\delta} \exp\left(\frac{2}{3}N\delta^3\right). \qquad (39)$$

Equation (39) can be also obtained from a Fokker–Planck equation, and it is universally applicable to a whole class of well-mixed single-population models that have a very strong Allee effect. This result also holds for a whole class of continuous stochastic systems [23].

2.2.2. Switching. The difference between switching and extinction is that, in the case of switching, the target state, corresponding to the second attracting point n_s, is non-absorbing. As explained in the previous section, the mean time to escape in scenario B does not depend on the exact nature of the target state. Still, for the sake of completeness, we will briefly show how to evaluate the mean switching time using, as an example, a genetic switch.
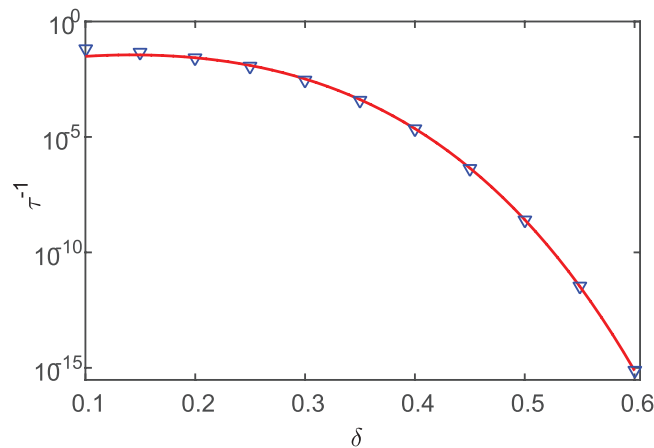
Figure 7. The mean extinction rate τ^{-1} versus δ on a semi-log scale. Line: theoretical result (38). Symbols: results of a numerical solution of the master equation set for $N = 200$ and $\mu = 1$. Reprinted figure with permission from [35], copyright 2015 by the American Physical Society.

We will consider a gene regulatory network that displays a deterministically bistable behavior. Gene regulatory networks are responsible for regulating the production of proteins. During gene expression a DNA segment—the gene—is transcribed into an mRNA molecule which is then translated into a protein. This process is often regulated via transcription factors (which are also proteins) that can bind to the DNA promoter site and affect the mRNA transcription rate, and thereby the protein translation rate. Here control of the mRNA transcription is done by either recruiting or blocking RNA polymerase—an enzyme that performs the transcription of genetic information from DNA to RNA [88].

In some cases the dynamics of the transcription factor, that controls a specific protein, are strongly affected by the protein itself, so there is feedback. A *positive* feedback can give rise to a bistable behavior, which has been shown to occur e.g. in the *lac operon* circuit [89, 90] and also in the context of competence in *Bacillus subtilis* [91]. Bistability can also occur if two different proteins *negatively* regulate each other, see e.g. [45, 88, 92–95]. Such feedback-based genetic switches are abundant in cell biology. They regulate diverse decision-making processes such as microbial environmental adaptation, developmental pathways and nutrient homeostasis, see [89] and references therein. In such systems it has been shown that the lifetime of these different gene-expression states (phenotypes) is determined by stochastic fluctuations of mRNA and proteins during gene expression that can yield spontaneous switching, even in the absence of a deterministic signal [43, 96].

We will consider a simple genetic switch in which the protein of interest positively regulates the transcription of the mRNA molecule that is responsible for its own production, giving rise to a positive feedback loop [43, 51, 90, 96]. In many situations the lifetime of the mRNA is short compared to that of the protein [97]. Here one can make the simplifying assumption that the mRNA species instantaneously equilibrates, and the switching time depends solely on the protein dynamics. (We will revisit this assumption in section 7.1.7 where we will explicitly consider the mRNA dynamics as well.) This gives rise to the so-called 'self-regulating gene' which is a protein-only model. In this model the protein's production rate is given by a function $f(n)$ which depends on the current protein number n, while degradation, mainly due to cell division, occurs at a rate n [50]. Here time is measured in cell cycle units.

14

When positive feedback is at play, the function $f(n)$ increases with a growing protein number. A widely used positive-feedback function is the Hill function [89, 98]

$$f(n) = a_0 + (a_1 - a_0) \frac{n^h}{n^h + n_0^h}, \tag{40}$$

where h is the Hill exponent, a_0 is the baseline production rate, n_0 is its midpoint, and the production rate saturates at a_1 at $n \to \infty$. For $h \geqslant 2$ the deterministic rate equation for the mean number of proteins,

$$\dot{n} = f(n) - n \tag{41}$$

can give rise to bistability. In the bistable case, the rate equation admits (at least) three fixed points: a lower attracting point n_*, an intermediate repelling point n_u, and a higher attracting point at n_s. In cellular biology the attracting fixed points represent 'off' and 'on' phenotypes, in which the cell produces a low/high number of proteins, respectively. When stochasticity is accounted for, one observes noise-driven switching between the two phenotypic states, each of which now becoming metastable.

The stochastic description of the self-regulating gene is given in terms of the master equation

$$\frac{dP_n(t)}{dt} = f(n-1)P_{n-1}(t) + (n+1)P_{n+1}(t) - [f(n) + n]P_n(t). \tag{42}$$

What is the mean switching time (MST) to the on state $n = n_s$ when starting from the vicinity of the off state $n = n_*$? As usual, we first determine the QSD π_n via the ansatz $P_{n<n_u}(t) = \pi_n e^{-t/\tau}$, where τ is the MST, and employ the WKB ansatz (10) [33, 50]. In the leading order, this gives rise to a Hamilton–Jacobi equation $H(q, dS/dq) = 0$, where

$$H(q, p) = \tilde{f}(q) \left(e^p - 1 \right) + q \left(e^{-p} - 1 \right) \tag{43}$$

is the Hamiltonian, $\tilde{f}(q) = f(n)/N$, $q = n/N$ is the rescaled protein concentration, and N is the typical protein population size. As a result, the MST in the leading order, $\tau \sim e^{N \Delta S}$, is given by

$$\Delta S = \int_{q_*}^{q_u} \ln[q/\tilde{f}(q)] dq. \tag{44}$$

As the target state n_s is *not* absorbing, there is actually a backward probability current towards n_*. However, when starting from the vicinity of the off state $n = n_*$, and at intermediate times $t_r \ll t \ll \tau$, the backward current is exponentially small compared to the forward current, and therefore, does not affect the MST [33, 35].

As discussed above, the subleading-order corrections to the MST can be calculated by matching the fast-mode WKB solution with a boundary-layer solution valid in the vicinity of n_u, and subsequently, matching the boundary-layer solution to the slow-mode WKB solution. This allows the calculation of the probability current through the repelling point and subsequently the MST [33, 35].

3. Extinction conditioned on non-establishment

In this section we consider a failure of establishment. The question we are asking is the following: What is the probability that a population, with initial population size $n_0 \gg 1$, goes extinct *before* it gets established at an attracting fixed point n_*? This probability is exponentially small

and amenable to a WKB treatment. Here too there are two extinction scenarios: A and B. We will consider scenario A in some detail on the example of the Verhulst model with $W_{\pm 1}(n)$ given by equation (7). The repelling fixed point $n = 0$ is an absorbing state of our stochastic process. In order to answer the question we have posed, we declare the attracting fixed point $n = n_*$ absorbing, and find the probability of reaching $n = 0$ rather than $n = n_*$. Let us denote by $\mathcal{P}_n$ this conditional extinction probability starting from n individuals. We can write the following recursive equation [58]:

$$\mathcal{P}_n = \frac{W_{+1}(n)}{W_{+1}(n) + W_{-1}(n)} \mathcal{P}_{n+1} + \frac{W_{-1}(n)}{W_{+1}(n) + W_{-1}(n)} \mathcal{P}_{n-1}. \tag{45}$$

This equation follows from the observation that the probability of extinction starting from n individuals equals the sum of the probability of extinction starting from $n + 1$ individuals multiplied by the probability to reach state $n + 1$ and the probability of extinction starting from $n - 1$ individuals multiplied by the probability to reach state $n - 1$ [58]. The boundary conditions for $\mathcal{P}_n$ are $\mathcal{P}_0 = 1$ and $\mathcal{P}_{n_*} = 0$. Equation (45) can be rewritten as

$$W_{+1}(n)R_n - W_{-1}(n)R_{n-1} = 0, \tag{46}$$

where $R_n = \mathcal{P}_{n+1} - \mathcal{P}_n$. The ensuing single-step problem is in fact exactly solvable, see e.g. [58]. This is not so, however, in more complicated situations, where multi-step transitions are present. Therefore, we employ a WKB approximation and sketch the leading-order WKB solution. The WKB ansatz is $R_n \equiv R(q) = A_1 e^{-N\mathcal{S}(q)}$, $q = n/N$, and A_1 is a constant. Plugging it into equation (46) [36, 37] we arrive, in the leading order, at the problem of finding the non-trivial zero energy trajectory, generated by the Hamiltonian

$$H(q, p) = w_{+1}(q) - w_{-1}(q)e^p, \tag{47}$$

where $w_{\pm}(q) = W_{\pm}(n)/N$, and $p = \mathcal{S}'(q)$ is the momentum. The solution is $p(q) = \ln[w_{+1}(q)/w_{-1}(q)]$, and therefore $R(q) = A_1 e^{-N\mathcal{S}(q)}$, where

$$\mathcal{S}(q) = \int_q^0 p(q) \mathrm{d}q = -q + q \ln\left(\frac{1}{B} + q\right) + \frac{1}{B} \ln(1 + Bq). \tag{48}$$

The approximation we have made is valid at $n \gg 1$, or $q \gg N^{-1}$, and $N \gg 1$. As a result, we obtain

$$\mathcal{P}_n = A_2 + \sum_{m=0}^{n-1} R_m \simeq A_2 + NA_1 \int_0^q R(q) \mathrm{d}q, \tag{49}$$

where we have replaced the sum by an integral. A_1 and A_2 are constants which can be found using the boundary conditions $\mathcal{P}_0 = 1$ and $\mathcal{P}_{n_*} = 0$, and we obtain

$$\mathcal{P}_n \simeq \frac{\int_q^{q_*} e^{-N\mathcal{S}(q')} \mathrm{d}q'}{\int_0^{q_*} e^{-N\mathcal{S}(q')} \mathrm{d}q'}. \tag{50}$$

As $N \gg 1$, the integrals in the numerator and denominator can be evaluated by Taylor-expanding $\mathcal{S}(q')$ around the lower boundaries q and 0, respectively. Then, putting $n = n_0 < n_*$, we arrive at the final result [36, 37]

$$\mathcal{P}_{n_0} \simeq C(q_0)e^{-N\mathcal{S}(q_0)}, \tag{51}$$

where $q_0 = n_0/N$, $C(q_0) = A_3 \mathcal{S}'(0)/\mathcal{S}'(q_0)$, and A_3 is an additional pre-factor that we will not present here. It comes from two sources. The first is the subleading WKB term $\mathcal{S}_1(q)$. The

second is the boundary correction term in the Euler–Maclaurin formula of replacing a sum by an integral, that we neglected in equation (49).

The optimal path to extinction conditioned on non-establishment can be also found by solving a modified problem where a *reflecting wall* is put at $n = n_0 < n_*$ so that the population cannot have more than n_0 individuals. In the latter case the population *cannot* reach the metastable state at n_*. The pre-exponential factors in these two problems are, however, different.

In scenario B there are three fixed points of the deterministic theory: the attracting points $n = 0$ and $n = n_*$ and a repelling point $n = n_u$. Starting from $n = n_0 > n_u$, one is interested in the probability of extinction prior to reaching the established state at n_*. By similar arguments, the leading-order WKB result is

$$\mathcal{P}_{n_0} \simeq C(q_0)e^{-N[\mathcal{S}(q_u)-\mathcal{S}(q_0)]}, \tag{52}$$

where $C(q_0)$ is a preexponential factor which can be calculated.

4. Bursty reactions

Sometimes production or death of individuals occurs in 'bursts' of random size. This happens, for example, in living cells. When the life-time of mRNA molecules is short compared to the cell cycle, proteins are synthesized in geometrically-distributed bursts, see e.g. [97]. This fact alters protein statistics and may drastically decrease switching times between different phenotypic states [43]. Additional examples include bursty viral production from infected cells, see e.g. [99], and variations in the number of offspring in animals, which has been shown to decrease the extinction risk [100].

We will illustrate the problem of bursty reactions by considering a modification of the Verhulst model, presented above. (A similar approach can be taken to treat the problem of bursty influx, or arrival in groups, see [69].) The microscopic dynamics are defined by the following ensemble of reactions and their corresponding rates [70]

$$A \xrightarrow{BnD(k)/\langle k \rangle} A + kA, \quad A \xrightarrow{n+Bn^2/N} \emptyset. \tag{53}$$

Here, $N \gg 1$ is the typical population size, $B \gtrsim 1$ is the average reproduction rate, and $k = 0, 1, 2, \ldots$ is the offspring number per birth event, which is sampled from a normalized burst size distribution $D(k)$ with the first and second moments denoted by $\langle k \rangle$, and $\langle k^2 \rangle$, respectively.

In this case the master equation reads

$$\dot{P}_n = \frac{B}{\langle k \rangle} \left[\sum_{k=0}^{n-1} D(k)(n-k)P_{n-k} - \sum_{k=0}^{\infty} D(k)nP_n \right]$$
$$+ (n+1)\left[1 + \frac{B(n+1)}{N}\right]P_{n+1} - n\left[1 + \frac{Bn}{N}\right]P_n, \tag{54}$$

whereas the deterministic rate equation coincides with equation (9). The long-lived metastable state is peaked about the attracting fixed point $n_* = N(1 - 1/B)$, and our aim is to calculate the MTE. The WKB machinery leads to $H = 0$, with Hamiltonian [70]

$$H(q,p) = (1 - e^{-p})Bq\left[e^p F(p) - 1/B - q\right], \tag{55}$$

where

$$F(p) = \frac{\sum_{k=0}^{\infty} e^{kp} D(k) - 1}{\langle k \rangle (e^p - 1)}. \tag{56}$$

When $D(k)$ is specified, the activation trajectory $p_a(q)$ can be found explicitly. Having found $p_a(q)$, one can show that the QSD's variance satisfies $\sigma^2 = N[1 + F'(0)]$, where $F'(0) = (1/2)(\langle k^2 \rangle / \langle k \rangle - 1)$. The MTE can be evaluated as [70]

$$\tau \sim e^{N\Delta S}, \quad \Delta S = \frac{p_a(0)}{B} - \int_0^{p_a(0)} e^p F(p) \mathrm{d}p. \tag{57}$$

Finding the subleading correction by the real-space WKB method would require matching of the subleading-order WKB solution with a recursive solution at $n = \mathcal{O}(1)$. It turns out, however, that the method of calculating the recursive solution by linearizing the reaction rates in the vicinity of $n = \mathcal{O}(1)$ breaks down in the presence of bursty reproduction [70]. Fortunately, the momentum-space approach comes to rescue. Here, as in section 2.1.1, we can match the bulk solution and the boundary-layer solution in their joint region of validity, see figure 6. This yields the MTE including the pre-exponential factor [70]:

$$\tau = \frac{1}{(B-1)(e^{-p_f} - 1)} \sqrt{\frac{2\pi}{N q_a'(p_f)}} e^{N\Delta S}, \tag{58}$$

where $q_a(p) = e^p F(p) - 1/B$ is the zero-energy solution of the Hamiltonian (55), p_f is the root of the equation $q_a(p_f) = 0$, and ΔS is given by equation (57). Figure 8 shows comparisons between the analytical and numerical results for τ for four different burst size distributions.

Close to the transcritical bifurcation of the deterministic rate equation, $\delta \equiv B - 1 \ll 1$, but still at $N\delta^2 \gg 1$, equation (58) simplifies to

$$\tau \simeq \frac{\sqrt{2\pi[1 + F'(0)]}}{\sqrt{N}\delta^2} \exp\left\{ \frac{N\delta^2}{2[1 + F'(0)]} \right\}. \tag{59}$$

Let us compare this result with equation (18) for the MTE of the 'standard' Verhulst model. As $F'(0) > 0$, equation (59) shows that bursty reproduction can reduce the MTE by an exponentially large factor [70].

Bursty deaths, where a random number of individuals dies (for example, due to competition), can be treated in a similar manner. Overall, bursty reactions can either increase or decrease the probability of rare events to occur compared to the non-bursty reactions [101].

5. Which stationary WKB method is better?

The stationary real-space and momentum-space WKB methods, in conjunction with additional approximations employing the same large parameter $N \gg 1$, yield accurate large-deviation results for a broad class of problems of single stochastic populations. Such accuracy is impossible to achieve with the more traditional van Kampen system size expansion which approximates the exact master equation by a Fokker–Planck equation and, as a result, usually holds only for typical fluctuations. How does the momentum-space WKB method compare with the real-space WKB method? From our experience, every problem with 'non-bursty' reactions, which includes a large parameter $N \gg 1$ and can be approximately solved with the momentum-space WKB, can also be approximately solved with the real-space WKB, and these approximate results coincide. For populations exhibiting escape scenario B (the Allee effect), the momentum space representation encounters significant difficulties [68].
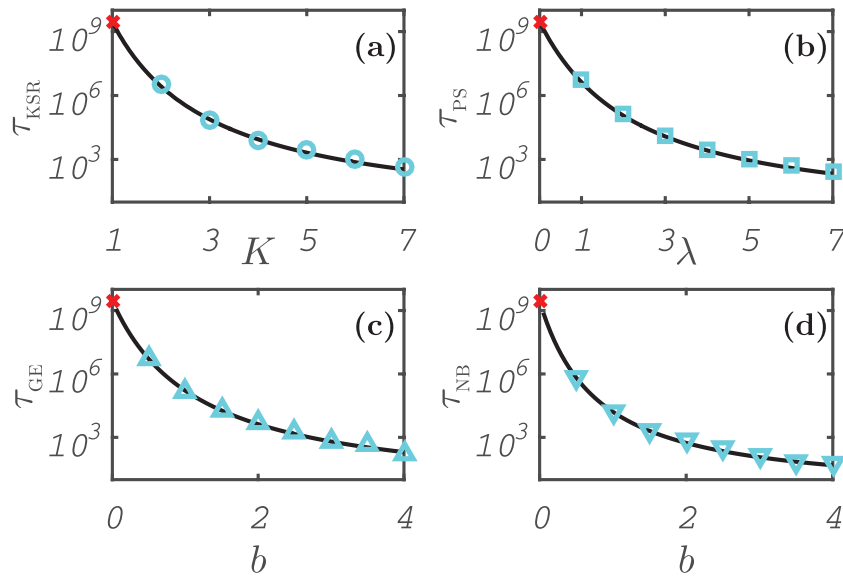
Figure 8. MTE for different burst size distributions as a function of their characteristic parameter: (a) K-step reproduction (KSR) $D(k) = \delta_{k,K}$; (b) Poisson (PS) $D(k) = \lambda^k e^{-\lambda}/k!$; (c) geometric (GE) $D(k) = (1/(1+b))(b/(1+b))^k$; (d) negative-binomial (NB) $D(k) = (k+a-1)!/(k!(a-1)!)(1/(1+b))^a (b/(1+b))^k$. The solid lines denote theoretical results given by equation (58), the ($\circ$), ($\square$), ($\triangle$) and ($\triangledown$) markers denote results of Monte-Carlo simulations, while the (x) marker denotes the theoretical value for the single-step case with $D(k) = \delta_{k,1}$. Parameters are $a = 2$, $B = 2$, and $N = 150$. Reproduced from [70], © SISSA Medialab Srl. Reproduced by permission of IOP Publishing. All rights reserved.

On the other hand, for problems exhibiting bursty reactions, the momentum-space approach is, as of today, the only method that enables one to calculate the preexponential corrections, see section 4 and [69, 70]. In this case even the leading-order WKB calculations are more conveniently done with the momentum-space approach. Here the master equation, which contains an infinite sum corresponding to the infinite reaction ensemble, transforms into a single evolution equation for the probability generating function, $G(p, t)$, where the bursty nature of the reactions is manifested by a nontrivial p-dependence of this equation. By contrast, the Hamiltonian obtained via the real-space approach contains an infinite number of terms, which makes theory more complicated [69, 70].

6. Extinction due to demographic noise and environmental variations

6.1. Periodic modulation of the environment

A periodic rate modulation can greatly reduce the MTE. The authors of [64, 66] studied this effect on the prototypical example of the branching-annihilation process $A \to 2A$ and $2A \to \emptyset$. The transition $2A \to \emptyset$ can be viewed as an extreme variant of competition when two competitors are both killed upon encounter. The transition rate coefficients of the branching and annihilation processes are λ and μ_0, respectively. Let λ oscillate with time, $\lambda = \lambda_0(1 + \epsilon \cos \omega t)$, so as to describe e.g. seasonal effects. The master equation reads, for $n \geqslant 1$,

$$\frac{dP_n(t)}{dt} = \frac{\mu_0}{2}\left[(n+2)(n+1)P_{n+2}(t) - n(n-1)P_n(t)\right] + \lambda(t)\left[(n-1)P_{n-1}(t) - nP_n(t)\right].$$
(60)

Using this equation, one obtains the following exact evolution equation for $G(p,t)$:

$$\frac{\partial G}{\partial t} = \frac{\mu_0}{2}(1-p^2)\frac{\partial^2 G}{\partial p^2} + \lambda(t)p(p-1)\frac{\partial G}{\partial p}.$$
(61)

In the absence of modulation, $\epsilon = 0$, the MTE is the following [30, 63, 102]:

$$\tau \simeq \frac{2\sqrt{\pi}\mu^{1/2}}{\lambda_0^{3/2}}e^{\frac{2\lambda_0}{\mu}(1-\ln 2)} = \frac{2\sqrt{\pi}}{\mu N^{3/2}}e^{2N(1-\ln 2)},$$
(62)

where $N = \lambda_0/\mu \gg 1$. This result (which includes an important pre-exponential factor) was obtained by three different methods, among them the real-space WKB method in the leading and subleading WKB order [30]. For the time-periodic λ, such a high accuracy can only be achieved in the limit when the modulation frequency is very low, see below. In other regimes one settles for the leading-order WKB accuracy. Let us use the non-stationary WKB ansatz in the momentum space $G(p,t) = \exp[-S(p,t)]$ (here and in section 6.2 we absorb N in the definition of S). Defining $q = \partial S/\partial p$, neglecting $\partial^2 S/\partial p^2$ and shifting the momentum, $p \to p-1$, we arrive at an effective one-dimensional classical mechanics with the Hamiltonian

$$H(q,p) = \left[\lambda(t)(1+p) - \frac{\mu}{2}(2+p)q\right]qp$$
(63)

that explicitly depends on time. This mechanical problem is in general not solvable analytically. Two existing numerical algorithms will be briefly reviewed in section 7.1.1. Analytical progress requires additional approximations, based on additional small parameters. Assaf *et al* [66] used three such approximations.

6.1.1. Weak modulation: linear theory. At sufficiently small, but not too small, ϵ, the relatively small, but important correction to the action is linear with respect to ϵ, and can be calculated perturbatively within the WKB approximation. We split the Hamiltonian into two terms:

$$H(q,p,t) = H_0(q,p) + \epsilon H_1(q,p,t),$$
(64)

where

$$H_0(q,p) = \lambda_0(1+p)pq - \frac{\mu_0}{2}(2+p)pq^2,$$
(65)

and

$$H_1(q,p,t) = \lambda_0(1+p)pq\cos\omega t.$$
(66)

For the unperturbed Hamiltonian $H_0(q,p)$ the 'extinction instanton'—the heteroclinic trajectory MF of figure 9—is [63]

$$q = q_0(p) = \frac{2N(1+p)}{(2+p)}.$$
(67)

One can find explicit formulas for $q(t)$ and $p(t)$:

$$q_0(t-t_0) = \frac{2N}{2+e^{\lambda(t-t_0)}}, \quad p_0(t-t_0) = -\frac{1}{1+e^{-\lambda(t-t_0)}},$$
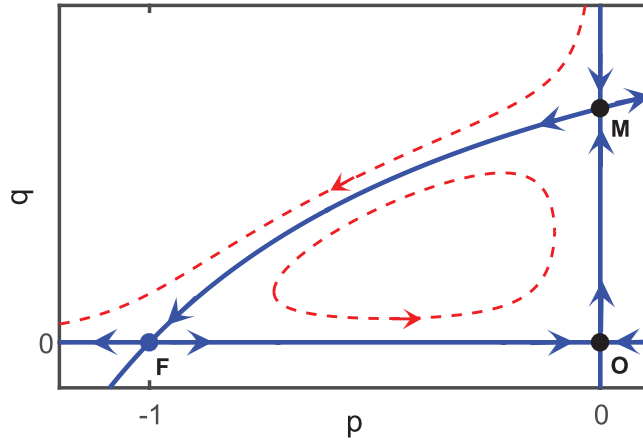(68)

Figure 9. The phase plane of the unperturbed Hamiltonian (65). The solid lines denote zero-energy trajectories. Reprinted figure with permission from [66], copyright 2015 by the American Physical Society.

where t_0 is an arbitrary time shift. Using equation (67), we obtain the unperturbed action S_0 along the instanton,

$$S_0 = -\int_0^{-1} q_0(p)\,\mathrm{d}p = 2N(1 - \ln 2) \tag{69}$$

which yields the expression in the exponent of equation (62).

A small correction to the action ΔS can be evaluated from [66]

$$\Delta S = \min_{t_0}\left\{-\varepsilon\int_{-\infty}^{\infty} H_1\left[q_0(t-t_0), p_0(t-t_0), t\right]\,\mathrm{d}t\right\}, \tag{70}$$

where the integration is performed along the *unperturbed* instanton (68). After the integration over time and minimization over t_0, the perturbed action $S = S_0 + \Delta S$ satisfies

$$\frac{S}{S_0} = 1 + \frac{\Delta S}{S_0} \simeq 1 - \frac{\pi\varepsilon}{(1-\ln 2)\sinh(\pi\alpha)}\left\{[\sin(\alpha\ln 2) - \alpha]^2 + [\cos(\alpha\ln 2) - 1]^2\right\}^{1/2}, \tag{71}$$

where $\alpha = \omega/\lambda_0$. The maximum effect of the rate modulation is obtained at $\alpha \to 0$. Recall that the perturbed action yields (in the leading order in N) minus the natural logarithm of the MTE.

The minimization over t_0 reflects an important effect of synchronization between the optimal path of the population to extinction and the periodic rate modulation. The synchronization lifts the degeneracy of the unperturbed instanton with respect to an arbitrary time shift present in equation (68). Synchronization of this nature was also observed in large deviations of continuous stochastic systems subject to a parameter modulation [103, 104].

6.1.2. Fast modulation: the Kapitsa's pendulum. In the the high-frequency limit, $\omega \gg \lambda_0$, the system resembles the Kapitsa's pendulum [105]: a rigid pendulum with a rapidly vibrating pivot. Here one can also calculate an important correction to the unperturbed action. Because of the high modulation frequency, this correction is small even if ε is of order 1. To determine this correction Assaf *et al* [66] developed a Hamiltonian version of the asymptotic method for the Kapitsa's pendulum [105]. As a first step, one calculates small high-frequency corrections

to the unperturbed 'coordinate' and 'momentum' of the system. Using these corrections, one then constructs a canonical transformation which, by means of a time averaging (essentially, by rectifying the high-frequency component of the motion), transforms the original time-dependent Hamiltonian into an effective *time-independent* Hamiltonian. The effective Hamiltonian includes a correction coming from the rectified high-frequency perturbation. Finally, one finds the perturbed instanton, emerging from this effective time-independent Hamiltonian, and evaluates the action along this instanton. The action along the effective instanton turns out to be [66]

$$\frac{S}{S_0} \simeq 1 - \frac{K\epsilon^2}{\alpha^2}, \tag{72}$$

where

$$K = \frac{6\ln 2 - 49/12}{1 - \ln 2} = 0.2462\ldots. $$

6.1.3. Slow modulation: adiabatic approximation. When the modulation period $2\pi/\omega$ is much longer than the characteristic relaxation time of the system $t_r = 1/\lambda_0$ but much shorter than the expected MTE, the extinction probability as a function of time can be written, at long times, as [66]

$$P_0(t) \simeq 1 - e^{-\int_0^t r_{\text{ext}}(t)\,\mathrm{d}t}. \tag{73}$$

Here $r_{\text{ext}}(t) = r_{\text{ext}}[\lambda = \lambda(t)]$ is the *instantaneous value* of the slowly time-dependent extinction rate. It is obtained by replacing λ_0 in equation (62) by $\lambda(t) = \lambda_0(1 + \epsilon\cos\omega t)$. The *average*, or effective extinction rate $\bar{r}_{\text{ext}}$ during a sufficiently long time T is defined via the relation

$$P_0(T) = 1 - e^{-\bar{r}_{\text{ext}}T}. \tag{74}$$

Comparing equations (73) and (74), we obtain for our periodic rate modulation

$$\bar{r}_{\text{ext}} = \frac{2\pi}{\omega}\int_0^{2\pi/\omega} r_{\text{ext}}(t)\,\mathrm{d}t = \frac{N^{3/2}}{4\pi^{3/2}}\int_0^{2\pi}(1 + \varepsilon\cos\tau)^{3/2}e^{-S_0(1+\varepsilon\cos\tau)}\mathrm{d}\tau. \tag{75}$$

The MTE is equal, in the adiabatic approximation, to $1/\bar{r}_{\text{ext}}$. In [66] the integral over τ in equation (75) was evaluated analytically in two limits. For $|\epsilon|N \gg 1$, a saddle point evaluation yields

$$\bar{r}_{\text{ext}} = \frac{N(1 - |\epsilon|)^{3/2}}{4\pi\sqrt{|\epsilon|(1 - \ln 2)}}e^{-S_0(1-|\epsilon|)}. \tag{76}$$

The leading term $S_0(1 - |\varepsilon|)$ coincides with the zero-frequency limit of the linear theory, as can be seen from equation (71) for $\alpha \to 0$. The physical meaning of this term is evident: by virtue of the adiabatically slow rate modulation, the 'activation barrier' to extinction $S_0(1 - |\varepsilon|)$ is determined simply by the minimum value of $\lambda(t) = \lambda_0(1 + \varepsilon\cos\omega t)$ which is equal to $\lambda_0(1 - |\varepsilon|)$. It is not surprising, therefore, that the same leading order result also follows from the leading-order WKB theory [66]. However, equation (76) also includes an important prefactor missed by the leading-order WKB. Notice that this prefactor scales as N, not as $N^{3/2}$ as observed without modulation, see equation (62).

In the opposite limit, $|\varepsilon|N \ll 1$, the integral (75) can be calculated via a Taylor expansion of the integrand in εN [66].
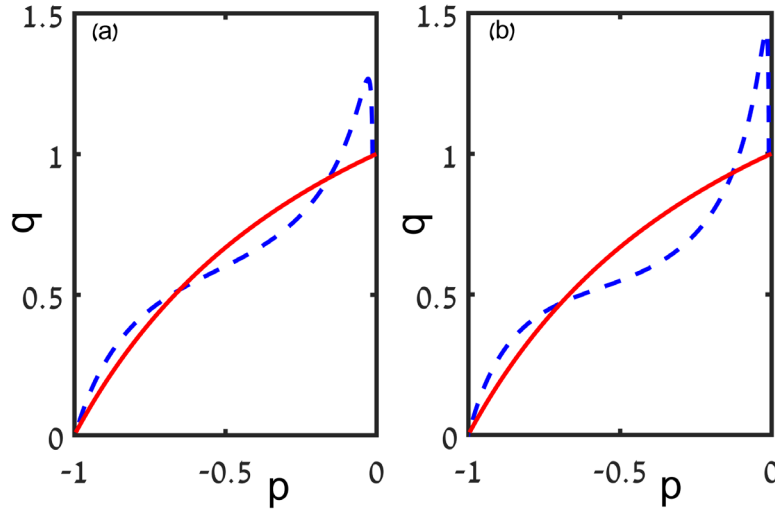
22

Figure 10. Projections on the (q, p) plane of instantons of the Hamiltonian (63), found numerically via the shooting method, for two sets of parameters (the dashed lines). Here q is measured in units of N, $\alpha = 1$, and $\varepsilon = 0.6$ such that $\mathcal{S}/N = 0.467$ in (a) and $\varepsilon = 0.9$ such that $\mathcal{S}/N = 0.398$ in (b). The solid line denotes the unperturbed instanton, whose unperturbed action is $\mathcal{S}_0/N = 2(1 - \ln 2) \simeq 0.614$. Reprinted figure with permission from [66], copyright 2015 by the American Physical Society.

The non-perturbative regime can only be studied numerically. The existing numerical methods are essentially the same as for multi-population problems, and we postpone their brief discussion until section 7.1.1. Here we only show, in figure 10, two examples of extinction instantons in this system, found numerically in [66]. A schematic 'phase diagram' of the system, on the plane $(\varepsilon, \alpha = \omega/\lambda_0)$, is shown in figure 11.

6.2. Catastrophic events

A catastrophic event in population dynamics can be modeled by a strong temporary decrease of the population birth rate, or increase of the death rate. How can one evaluate $P_0(t)$—the probability that the population goes extinct by time t—at the time when the catastrophic event is over, and the environmental conditions return to normal? As shown in [67], a catastrophe can lead to an exponentially large increase ΔP_0 in the extinction probability $P_0(t)$ compared to the extinction probability without the catastrophe, see figure 12. This increase ΔP_0 can be evaluated using the WKB approximation.

The authors of [67] considered the Verhulst model with the birth and death rates given by equation (7). The catastrophic event was modeled by introducing a time-dependent factor $f(t)$, so that $f(\pm\infty) = 1$, into the birth rate:

$$W_{+1}(n, t) = Bf(t)n. \tag{77}$$

The evolution equation for the probability generating function can be written as $\partial G/\partial t = \hat{\mathcal{H}}G$, with the operator

$$\hat{\mathcal{H}} = \frac{B}{N}(1 - p)\, p\, \frac{\partial^2}{\partial p^2} + (p - 1)\left[Bf(t)p - 1\right]\frac{\partial}{\partial p}. \tag{78}$$
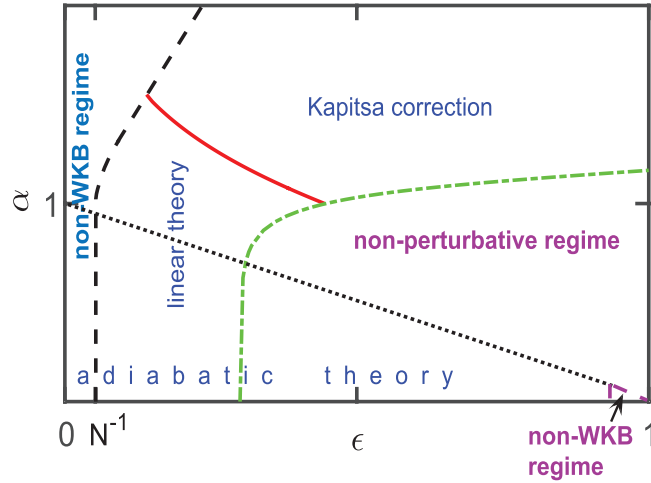
Figure 11. Schematic phase diagram of different extinction regimes in the parameter plane $(\epsilon, \alpha = \omega/\lambda_0)$ of the branching-annihilation process with modulation of the branching rate. Either the linear theory, or the hamiltonian Kapitza's method is applicable well to the left of and above the dash-dotted line, and well to the right of the dashed line. The position of the solid line is calculated in [66]. The adiabatic approximation holds well below the dotted line. The horizontal size of the non-eikonal (that is, non-WKB) regions at $\alpha \to 0$ is of order $1/N$, where $N = \lambda_0/\mu \gg 1$. There is no analytic theory as of today in the non-perturbative regime. Reprinted figure with permission from [66], copyright 2015 by the American Physical Society.

Similarly to the problem of rate modulation considered in the previous section, the non-stationary WKB ansatz in the momentum space leads, in the leading order, to a one-dimensional classical Hamiltonian flow with the time-dependent Hamiltonian

$$H(p,q,t) = p \left[-\frac{B}{N}(p+1)q + Bf(t)(p+1) - 1 \right] q, \tag{79}$$

where the new momentum p is shifted by 1, $p - 1 \to p$. Because of the explicit time dependence of the Hamiltonian, only numerical solution is in general available. One analytically solvable case is when the reproduction rate drops instantaneously (for simplicity, to zero) at $t = 0$ and recovers to the pre-catastrophe value, again instantaneously, after time T has elapsed:

$$f(t) = \begin{cases} 1 & \text{if } t < 0 \text{ or } t > T, \\ 0 & \text{if } 0 < t < T. \end{cases} \tag{80}$$

Because of the special shape of $f(t)$, there are now two distinct Hamiltonians to consider: the unperturbed Hamiltonian (equation (79) with $f(t) = 1$) before and after the catastrophe and the zero-birth-rate Hamiltonian during the catastrophe:

$$H_c(p,q) = -p \left[\frac{B}{N}(p+1)q + 1 \right] q. \tag{81}$$

Each of the two Hamiltonians is an integral of motion on the corresponding time interval. The instanton can be found by matching three separate trajectory segments: the pre-catastrophe, catastrophe and post-catastrophe segments. Figure 13 shows a projection of the instanton on the (p, q) plane. The instanton must exit, at $t = -\infty$, the deterministic fixed point M and enter,
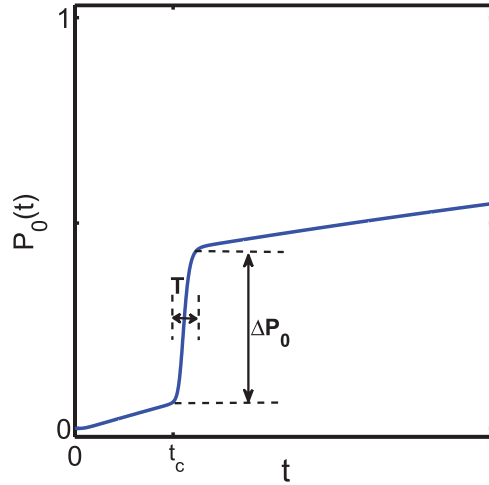
Figure 12. A schematic plot of the time-dependent extinction probability $P_0(t)$, showing the effect of a catastrophe with characteristic duration T. The increase of the extinction probability ΔP_0 can be evaluated using the WKB approximation. Reprinted figure with permission from [67], copyright 2015 by the American Physical Society.

at $t = \infty$, the fluctuational fixed point F. The matching conditions at times $t = 0$ and $t = T$ are provided by the continuity of the functions $q(t)$ and $p(t)$. The pre- and post-catastrophe segments must have a zero energy, $E = 0$, so they are parts of the original zero-energy trajectory of the unperturbed problem, $q_0(p) = N - N/[B(1 + p)]$. For the catastrophe segment, $q_c(p)$, however, the energy $E = E_c$ is non-zero and *a priori* unknown. It can be found from the demand that the duration of this segment be T [67].

The solution simplifies considerably at $B - 1 \ll 1$, that is close to the bifurcation point of the model. In this limit the pre- and post-catastrophe Hamiltonian reduces to the 'universal' Hamiltonian

$$H(p, q) = p\left(-\frac{q}{N} + p + B - 1\right)q, \tag{82}$$

introduced in [106]. Furthermore, the zero-birth-rate Hamiltonian during the catastrophe simplifies dramatically:

$$H_c(p, q) \simeq -pq. \tag{83}$$

After some algebra, see [67], one obtains

$$S(T) \simeq \frac{2S_0}{1 + e^T}, \tag{84}$$

where

$$S_0 \simeq \frac{N(B - 1)^2}{2}, \tag{85}$$

is the action of the 'universal' model in the absence of a catastrophe. The increase ΔP_0 of the extinction probability due to the catastrophe is $\sim\exp[-S(T)]$. As one can see, for a prolonged catastrophe, $T \gg 1$, the increase of the extinction probability becomes very large.
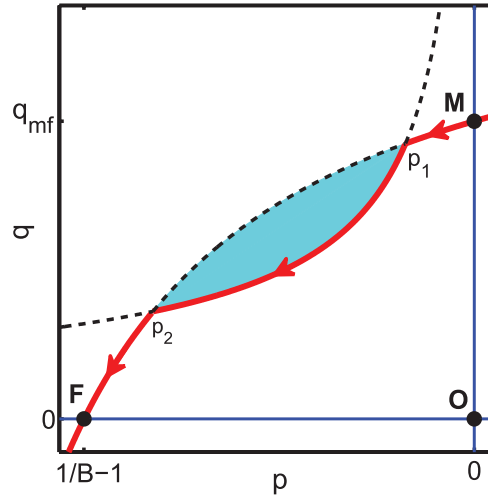
25

Figure 13. Projection on the (p, q) plane of the 'extinction instanton' (the thick solid line going from M to F) for the catastrophic event described by equations (77) and (80). Points p_1 and p_2 correspond to times $t = 0$ and $t = T$ where the catastrophic event begins and ends, respectively. At $t < 0$ and $t > T$ the instanton follows the unperturbed zero-energy heteroclinic trajectory $q = q_0(p)$, whereas at $0 < t < T$ it follows a trajectory with a nonzero energy which depends on the duration T of the catastrophe. The increase ΔP_0 of the extinction probability due to the catastrophe is $\sim \exp[-S(T)]$, where $S(T)$ is area of the shaded region. Reprinted figure with permission from [67], copyright 2015 by the American Physical Society.

6.3. Environmental noise

Multiple and concurrent environmental factors that affect the birth or death rates of populations can be modeled as a rate modulation by *noise*. Early population biology models assumed that the environmental noise is white, that is delta-correlated in time [107, 108]. More recent studies focused on the effect of temporal autocorrelation, or color, of the environmental noise on the population extinction, see e.g. [14] for a review. WKB theories of population extinction driven by an interplay of demographic and environmental noise in the absence of an Allee effect and in the presence of it were developed in [74] and [75], respectively. Kamenev *et al* [74] considered a symmetrized Verhulst model with the birth and death rates

$$\lambda_n = \frac{n}{2}(\mu + r - an), \qquad \mu_n = \frac{n}{2}(\mu - r + an), \quad r < \mu,$$

(86)

where there is no Allee effect. Levine and Meerson [75] considered the set of reactions $2A \rightarrow 3A$, $3A \rightarrow 2A$ and $A \rightarrow \emptyset$, that we dealt with in section 2.2.1, in the parameter region corresponding to a strong Allee effect. In both papers the environmental noise was introduced via a modulation of a parameter entering the reaction rates. In [74] the constant parameter r was replaced by $r - \xi(t)$, where $\xi(t)$ is a 'red' (that is, positively correlated) Gaussian random process—the Ornstein–Uhlenbeck noise [58]—with zero mean, variance $v \ll \mu^2$ and correlation time t_c. The same type of noise was adopted in [75]. Both papers assumed a regime of parameters close to the bifurcation (transcritical and saddle-node, respectively) of the corresponding deterministic equations. In these cases the master equation for the evolution of the population size distribution *without* environmental noise can be well approximated by a Fokker–Planck equation. The latter is equivalent to a Langevin equation, so that the environmental noise $\xi(t)$ can be conveniently introduced in the birth-death term [75]. Importantly, the Ornstein–Uhlenbeck noise $\xi(t)$ obeys a Langevin equation of its own,

$$\dot{\xi}(t) = -\frac{\xi(t)}{t_c} + \sqrt{\frac{2v}{t_c}}\,\eta(t), \tag{87}$$

where $\eta(t)$ is a Gaussian *white* noise with zero mean and $\langle\eta(t_1)\eta(t_2)\rangle = \delta(t_1 - t_2)$. The resulting two-component Langevin equation leads to a two-dimensional Fokker–Planck equation for the joint probability $P(n, \xi, t)$ of observing a certain population size n and a certain value of the noise ξ at time t. Applying the WKB approximation to this two-dimensional Fokker–Planck equation, one arrives at a two-dimensional effective mechanical problem where, in the context of population extinction, one should again look for an instanton solution. In general, this two-dimensional problem can only be solved numerically. Perturbative analytical solutions can be obtained in the limits of short-correlated, long-correlated and weak environmental noise. The details can be found in [74, 75]. Here we will only summarize the main results of these two works. In both cases the environmental noise causes an exponentially large reduction of the MTE. At fixed variance of the environmental noise, positive correlations quicken extinction.

In the absence of the Allee effect the population-size dependence of the MTE changes from exponential without noise to a power law for a strong short-correlated noise and to almost no dependence for a long-correlated noise [74]. The power-law dependence of the MTE on the population size for a strong white environmental noise in the absence of the Allee effect was known for quite a while [107, 108]. The WKB theory also gives the optimal environmental fluctuation—a non-random function $\xi(t)$—that mostly contributes to the population extinction. The optimal fluctuation looks quite differently in different regimes, see figure 14. For a long-correlated environmental noise the effective birth rate slowly goes down with time, and then slowly recovers to its original value. In this case most of the 'extinction current' (the probability current to the absorbing state) is observed around the time when the effective birth rate is at a minimum, as to be expected on the physical grounds.

For short-correlated noise the optimal fluctuation is less intuitive. Surprisingly, it has a form of a catastrophe, when the effective birth rate abruptly drops on a certain time interval (which the theory predicts), stays almost constant and then returns, again abruptly, to its original value [74]. By predicting the optimal realizations of the environmental noise in different regimes, the WKB method provides, in addition to the MTE, an instructive visualization and a better understanding of the rare events where the environmental and demographic noises 'coconspire' to bring the population to extinction most effectively.

In the presence of a strong Allee effect the population-size dependence of the MTE changes from exponential for weak environmental noise to (approximately) no dependence for strong noise, implying a greatly increased extinction risk [75]. The exponential-to-power-law crossover of the MTE versus the population size, observed in the absence of the Allee effect, does not happen in the presence of the Allee effect. Here again, at fixed variance of the environmental noise, the noise correlations quicken extinction. Furthermore, the results of [75] imply that, at fixed variance, the MTE has a minimum at an intermediate value of the rescaled correlation time of the noise. If one instead fixes the noise intensity (the product of the variance and correlation time), the MTE grows monotonically with the rescaled correlation time and reaches a plateau at long correlation times.

Apart from population biology, environmental, or extrinsic noise may play an important role in gene regulation [109]. The extrinsic noise results from a multitude of factors such as variation in the cells' number of ribosomes, transcription factors and polymerases, fluctuations in the cell division time, as well as environmental fluctuations. The effect of extrinsic noise on the MST of a model gene regulatory network with positive feedback, see section 2.2.2, was studied in [50]. As in [74, 75], the authors assumed the Ornstein–Uhlenbeck extrinsic noise and
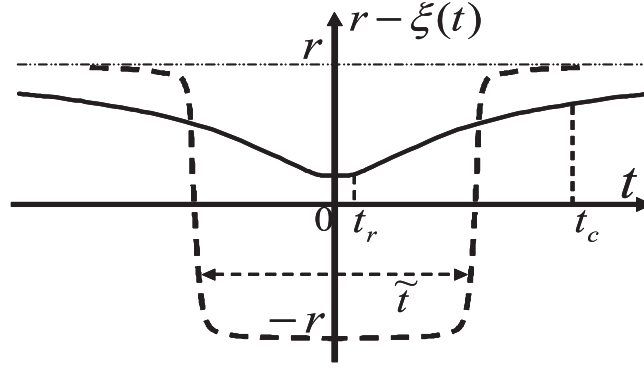
Figure 14. Optimal realizations of the environmental noise in the limit of long (the solid line) and short (the dashed line) noise correlations. $t_r = 1/r$ is the characteristic relaxation time of the deterministic theory. The duration of the 'catastrophe' for the short-correlated noise is $\tilde{t} \simeq t_r \ln(Kvt_c/\mu)$, where $K = r/a$. Reprinted figure with permission from [74], copyright 2015 by the American Physical Society.

worked close to the saddle-node bifurcation of the deterministic model. They derived a two-dimensional Fokker–Planck equation for the joint probability $P(n, \xi, t)$ of observing a protein copy-number n and extrinsic noise magnitude ξ at time t. They also observed that the extrinsic noise causes an exponential decrease of the MST between metastable states. In agreement with [75], they found that, at fixed variance of the extrinsic noise, there is an optimal correlation time, τ_c, for which the switching occurs at the maximum rate. A simple calculation, comparing the relative contributions of the short- and long-correlated extrinsic noise to the MST, shows that τ_c is inversely proportional to the extrinsic noise variance [50], see figure 15.

The treatment of environmental noise in [50, 74, 75] can be generalized to a non-gaussian noise, as well as to escape from a metastable state when the system is not necessarily close to bifurcation. For a non-gaussian noise the mean, variance and correlation time do not define the environmental noise uniquely. For gene regulatory networks one often observes reaction rates that are distributed according to a negative binomial distribution or a gamma distribution. From a theoretical viewpoint, such distributions have an advantage over the Gaussian distribution, as for the former distributions the reaction rates cannot reach a non-positive value, while the Gaussian distribution must be trimmed [53]. As an example, consider a gene regulatory network which is responsible for production of a protein of interest, and where the extrinsic variable is an auxiliary protein that affects the degradation rate of the protein of interest. Such an extrinsic noise variable ξ with, say, a negative-binomial distribution, mean $\langle \xi \rangle = 1$, and variance σ_{ex}^2 can be generated by defining a discrete variable $k = K\xi$, where $K \gg 1$ is a large integer, and using the following master equation

$$\dot{P}_{m,k}(t) = \frac{\alpha}{\tau_c}(P_{m-1,k} - P_{m,k}) + \frac{\omega}{\tau_c}[(m+1)P_{m+1,k} - mP_{m,k}]$$
$$+ \frac{\omega\beta m}{\tau_c}(P_{m,k-1} - P_{m,k}) + \frac{1}{\tau_c}[(k+1)P_{m,k+1} - kP_{m,k}]. \tag{88}$$

Here, $P_{m,k}(t)$ is the probability to create m auxiliary mRNA molecules and k auxiliary proteins at time t. This master equation describes the stochastic dynamics of the auxiliary circuit which includes mRNA transcription at a rate α/τ_c, mRNA degradation at a rate ω/τ_c, protein translation at a rate $\omega\beta/\tau_c$ and protein degradation at a rate $1/\tau_c$ [96, 97]. In the limit of short-lived mRNA, such that $\omega \gg 1$, it can be shown that the stationary probability distribution of the auxiliary protein is [97]
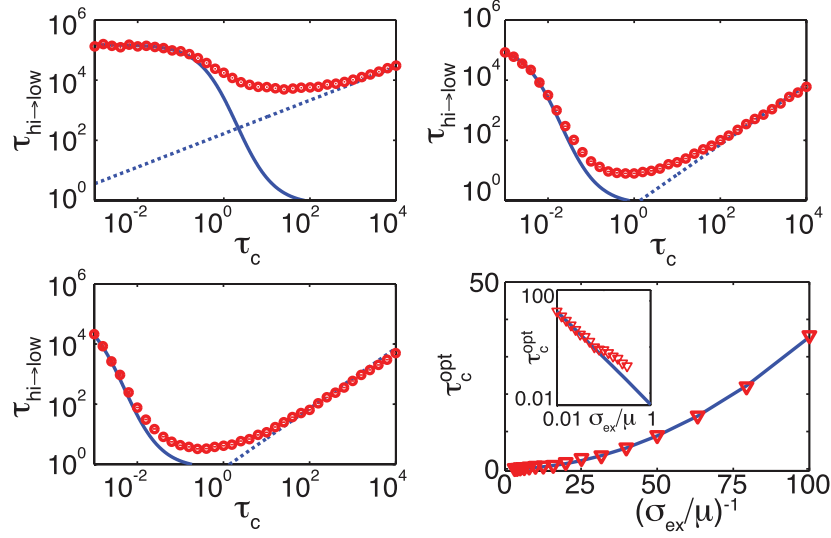
Figure 15. The MST versus the correlation time, τ_c, of the extrinsic noise for different variances of the extrinsic noise (a-c). $\sigma_{\mathrm{ex}}/\mu = 0.01$ (a), 0.1 (b) and 0.2 (c), where μ and σ_{ex} are the extrinsic noise's mean and standard deviation, respectively, while 'low' and 'hi' denote two long-lived metastable states, see section 2.2.2. The solid lines represent the WKB solution, the dashed lines represent an asymptotic analytical result in the *non*-WKB regime of strong adiabatic extrinsic noise. Here the perturbed action seizes to be large, and the MST is governed by the mean time it takes the Ornstein–Uhlenbeck process to reach such noise magnitude that enables almost instantaneous switching. Shown in (d) is the optimal correlation time as a function of the coefficient of variation, σ_{ex}/μ of the extrinsic noise. Reprinted figure with permission from [50], copyright 2015 by the American Physical Society.

$$P_k = \frac{\Gamma(\alpha + k)}{\Gamma(k+1)\Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^k \left(\frac{1}{\beta+1}\right)^{\alpha}, \tag{89}$$

where P_k is the probability to find k auxiliary proteins. The mean of this distribution is $\alpha\beta$, while its variance is $\alpha\beta(1+\beta)$. As a result, since $\xi = k/K$, by choosing $\alpha = 1/(\sigma_{\mathrm{ex}}^2 - 1/K)$ and $\beta = K\sigma_{\mathrm{ex}}^2 - 1$ we find that $\langle\xi\rangle = \langle k\rangle/K = 1$, and the variance of ξ becomes the variance of k over K^2 which equals σ_{ex}^2 as required. Finally, the correlation time of k or ξ is given by τ_c as required.

Using this definition of the extrinsic-noise variable, one can formulate a two-dimensional master equation for $P(n,k,t)$—the probability to find n proteins, and extrinsic noise magnitude $\xi = k/K$, at time t. This master equation can be analyzed by using a WKB approximation. The resulting Hamilton equations were solved for a short-correlated noise, by finding the optimal environmental fluctuation, and for a long-correlated noise, by adiabatic elimination of the fast variable [53, 110].

7. Large deviations in multi-population systems

When the demographic noise is weak, the quasi-stationary distribution of an established population is sharply peaked around an attractor of the corresponding deterministic theory. For example, for a generic two-population system this can be either an attracting fixed point or a stable limit cycle. The use of the WKB approximation for the analysis of multi-population

systems, initially residing around an attracting fixed point, was pioneered in [27] in the context of population switches, and in [31, 73] in the context of population extinction. As of today, there is a large number of works in these two directions, with applications ranging from population biology and ecology to dynamics of bacterial colonies to intracellular biochemistry. One way of applying the WKB ansatz is to the master equation describing the evolution of the probability $P_{m,n,\dots}(t)$ of observing $m, n, \dots$ individuals of each of the sub-populations at time t [27, 28, 31]. Alternatively, one can first derive an evolution equation for the probability generating function $G(p_m, p_n, \dots, t)$—a linear PDE with multiple arguments—and then apply the WKB ansatz [65]. In both cases the leading-order WKB approximation generates a multi-dimensional effective classical mechanical problem. Again, extinction of one or more of the sub-populations, or a switch from the vicinity of one attractor to another, is encoded in an instanton-like trajectory.

7.1. Established populations reside in a vicinity of a fixed point

Here we consider several examples from epidemiology, population dynamics and cell biochemistry. In each of these examples, in a WKB approximation, the problem of evaluating the mean time to extinction/switch boils down to finding an instanton in the phase space of the corresponding classical mechanics, and computing the action along it. As in one-population problems, the instanton is a zero-energy heteroclinic trajectory which exits a deterministic fixed point and approaches either a non-trivial 'fluctuational' fixed point or another deterministic fixed point. For a generic multiple-population system the instanton can only be found numerically, and we briefly review two existing numerical methods that have been proved useful. We also show how one can use additional small parameters to obtain approximate analytical solutions.

7.1.1. Extinction of endemic disease.
Our first example deals with 'endemic fadeout': spontaneous extinction of an infectious disease from a population after the disease has become endemic. The disease extinction ultimately occurs if the infectives recover, leave or die, while no new infectives are introduced into the susceptible population. The WKB approximation enables one to evaluate the MTE of the disease. Following [65], we will consider the SI (Susceptible-Infected) model of epidemiology with renewal of susceptibles [65, 73, 111–113]. In this model the population is divided into two dynamic sub-populations: susceptible (S) and infected (I). The set of processes and their rates are given in table 1.

The deterministic rate equations for the SI model can be written as

$$\dot{S} = \mu N - \mu S - (\beta/N)SI, \tag{90}$$

$$\dot{I} = -\mu_I I + (\beta/N)SI, \tag{91}$$

For a sufficiently high infection rate, $\beta > \mu_I$, there is a stable fixed point on the *S,I* plane,

$$\bar{S} = \frac{\mu_I}{\beta} N, \qquad \bar{I} = \frac{\mu(\beta - \mu_I)}{\beta \mu_I} N, \tag{92}$$

which describes an endemic state, and an unstable fixed point $\bar{S} = N$, $\bar{I} = 0$ (a saddle) which describes an infection-free steady-state population. At $\mu < 4\,(\beta - \mu_I)(\mu_I/\beta)^2$ the stable fixed point is a focus; for the opposite inequality it is a node. The inverse of the real part of the eigenvalues (for the focus), or the inverse of the smaller of the eigenvalues (for the node) yields the characteristic relaxation time t_r towards the 'endemic fixed point'.

Table 1. Stochastic SI model with renewal of susceptibles.

Process	Transition	Rate
Infection	$S \to S - 1, I \to I + 1$	$(\beta/N)SI$
Renewal of susceptible	$S \to S + 1$	μN
Removal of susceptible	$S \to S - 1$	μS
Removal of infected	$I \to I - 1$	$\mu_I I$

The master equation for the probability $P_{n,m}(t)$ of finding n susceptible and m infected individuals at time t has the form

$$
\begin{aligned}
\dot{P}_{n,m} &= \sum_{n',m'} M_{n,m;n',m'} P_{n',m'}(t) \\
&= \mu \left[N(P_{n-1,m} - P_{n,m}) + (n+1)P_{n+1,m} - nP_{n,m} \right] + \mu_I \left[(m+1)P_{n,m+1} - mP_{n,m} \right] \\
&\quad + (\beta/N) \left[(n+1)(m-1)P_{n+1,m-1} - nmP_{n,m} \right], \quad \text{for } m > 0,
\end{aligned}
\tag{93}
$$

and

$$
\dot{P}_{n,0} = \mu_I P_{n,1}.
\tag{94}
$$

Following [65], we use equation (93) to obtain an exact evolution equation for the probability generating function $G(p_S, p_I, t) = \sum_{n,m=0}^{\infty} p_S^n p_I^m P_{n,m}(t)$. This equation reads $\partial_t G = \hat{\mathcal{H}} G$ with the effective Hamiltonian operator

$$
\hat{\mathcal{H}} = \mu(p_S - 1)(N - \partial_{p_S}) - \mu_I(p_I - 1)\partial_{p_I} - (\beta/N)(p_S - p_I)p_I \partial^2_{p_S p_I}.
\tag{95}
$$

Using the WKB ansatz $G(p_S, p_I, t) = \exp[-\mathcal{S}(p_S, p_I, t)]$ and neglecting the second derivatives of the action $\mathcal{S}$ with respect to p_S and p_I, we arrive at a Hamilton–Jacobi equation $\partial_t \mathcal{S} + H = 0$ in the p-space with the classical Hamiltonian

$$
H(S, I, p_S, p_I) = \mu(p_S - 1)(N - S) - \mu_I(p_I - 1)I - (\beta/N)(p_S - p_I)p_I SI,
\tag{96}
$$

where $S = -\partial_{p_S}\mathcal{S}$ and $I = -\partial_{p_I}\mathcal{S}$. As H does not depend explicitly on time, it is an integral of motion: $H(S, I, p_S, p_I) = E = $ const. The deterministic trajectories, described by equations (90) and (91), lie in the zero-energy, $E = 0$, two-dimensional plane $p_S = p_I = 1$. We are interested in extinction starting from the long-lived, quasi-stationary endemic state. This requires an instanton: a non-deterministic activation trajectory. The stable fixed point (92) of the deterministic theory becomes a hyperbolic point $M_2 = [\bar{S}, \bar{I}, 1, 1]$ in the four-dimensional phase space with two stable and two unstable eigenvalues (the sum of which is zero) and respective eigenvectors. There are two more zero-energy fixed points that describe an infection-free population: the point $M_1 = [N, 0, 1, 1]$, which is also present in the deterministic description, and the 'fluctuational' fixed point $F = [N, 0, 1, \mu_I/\beta]$.

Let us sum up equation (94) over n. We obtain

$$
\frac{\mathrm{d}}{\mathrm{d}t} \sum_{n=0}^{\infty} P_{n,0} = \mu_I \sum_{n=0}^{\infty} P_{n,1}.
\tag{97}
$$

The left hand side in equation (97) describes the growth rate of the probability of disease extinction with time, whereas the right hand side is equal to μ_I times the probability of observing exactly one infected individual (at any number of susceptibles). At times much longer than the characteristic relaxation times of the deterministic theory, the probability distribution $P_{n,0}$ is sharply peaked close to the deterministic fixed point $(N, 0)$ corresponding to the

infection-free steady-state population. Essentially, by using the leading-order WKB approximation, one replaces the sums $\sum_n P_{n,0}$ and $\sum_n P_{n,1}$ by their maximum terms $P_{N,0}$ and $P_{N,1}$, respectively, for the purpose of calculating the probability flux to the infection-free absorbing state. Correspondingly, the instanton—the optimal path of the disease extinction—is a heteroclinic trajectory that exits, at $t = -\infty$, the 'endemic' fixed point M_2 along its two-dimensional unstable zero-energy manifold in the four-dimensional phase space and enters, at $t = \infty$, the fluctuational extinction point F, along its two-dimensional stable manifold [31, 65].

Up to a pre-exponent, the MTE of the disease is $\tau \propto \exp(\mathcal{S}_0)$, where

$$\mathcal{S}_0 = \int_{-\infty}^{\infty} (p_S \dot{S} + p_I \dot{I}) \, \mathrm{d}t, \tag{98}$$

and the integration is performed along the instanton going from M_2 to F. Following [65] we introduce new 'coordinates' $x = S/N - 1$ and $y = I/N$, time $\tilde{t} = \mu t$, new momenta $p_{x,y} = p_{S,I} - 1$ and the bifurcation parameter $\delta = 1 - \mu_I/\beta$, $0 < \delta < 1$. The action (98) can now be rewritten as $\mathcal{S}_0 = N\sigma$, where $\sigma(K, \delta)$ is the action along the proper instanton of the rescaled Hamiltonian

$$\tilde{H} = -p_x x - K\left[(1-\delta)p_y + (p_x - p_y)(p_y + 1)(x+1)\right]y \tag{99}$$

and $K = \beta/\mu > 1$. The fixed points M_1, M_2 and F become

$$[0, 0, 0, 0], \quad \left[-\delta, \ \frac{\delta}{K(1-\delta)}, 0, 0\right] \quad \text{and} \quad [0, 0, 0, -\delta],$$

respectively.

The endemic extinction instanton can be found numerically: either by the shooting method [28, 31, 65], or by the iteration method [114]. The shooting method works as follows. One linearizes the Hamilton equations near the 'endemic' fixed point and finds the two unstable eigenvectors. The instanton is then found by looking for the correct linear combination of the two unstable eigenvectors of a fixed (and very small) norm. An example of an instanton, found by the shooting method in [65], is shown in figure 16. In this case $4K\delta(1-\delta)^2 > 1$, and the endemic fixed point M_2 is a focus. For comparison, figure 16(a) also shows the deterministic heteroclinic trajectory ($p_x = p_y = 0$) connecting the fixed points M_1 and M_2. It describes, in deterministic terms, an epidemic outbreak following an introduction of a small number of infectives into a steady-state susceptible population. Deterministically, this outbreak always leads to an endemic state. In contrast to equilibrium systems, the x, y projection of the optimal path of a large fluctuation is different from the corresponding time-reversed relaxation path. As one can see, the difference between the 'activation spiral' and the time-reversed 'relaxation spiral' in this example is striking.

The shooting method usually works well for well-mixed two-population systems (and for one-population systems with time-dependent rates, see figure 10 in section 6). For three-population systems it becomes much less convenient, as now one needs to do shooting with respect to two parameters. The shooting method of course becomes impractical for spatially explicit systems (which, after discretization, represent multi-dimensional systems with a very large number of 'dimensions'). Here the back-and-forth iteration method comes to rescue. The iteration method was suggested by Chernykh and Stepanov [114] in the context of the instanton theory of the 'Burgers turbulence' [115, 116]. Then, starting from [60], it was adopted for population dynamics and for a host of other stochastic systems where a WKB approximation is used, and the ensuing Hamilton's equations need to be solved. The back-and-forth iteration method is applicable if the boundary conditions in time involve the knowledge of x and y at the initial time and p_x and p_y at the final time. For zero-energy instantons the initial time is at $-\infty$, and the final time is at $+\infty$. The iterations proceed as follows. One solves the Hamilton
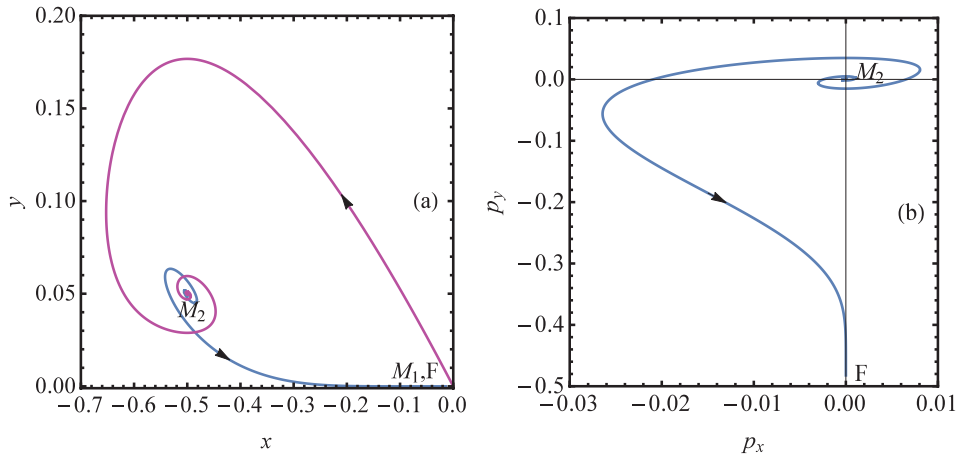
Figure 16. (a) The x, y projection of the four-dimensional optimal path in the phase space of the SI model with population turnover (the lower curve) and the deterministic trajectory ($p_x = p_y = 0$) describing an epidemic outbreak (the upper curve). (b) The p_x, p_y projection of the optimal path. $x = S/N - 1$, $y = I/N$; $K = 20$ and $\delta = 0.5$.

equations for $\dot{x}$ and $\dot{y}$ forward in time (with the initial conditions for x and y and with $p_x(t)$ and $p_y(t)$ from the previous iteration), and equations for $\dot{p}_x$ and $\dot{p}_y$ backward in time (with the 'initial' conditions for p_x and p_y and with $x(t)$ and $y(t)$ from the previous iteration). At the very first iteration reasonable 'seed functions' for $p_x(t)$ and $p_y(t)$ are used. A recent review [117] presents a detailed discussion of the Chernykh-Stepanov algorithm and its modifications and extensions in the context of hydrodynamic turbulence.

Let us return to the endemic fadeout in the SI model. At $\delta \ll 1$—close to the bifurcation point corresponding to the emergence of an endemic state—the system exhibits time-scale separation, and the optimal path to extinction and the action can be calculated analytically [31, 65]. Introducing rescaled variables $q_1 = x/\delta$, $q_2 = yK/\delta$, $p_1 = p_x/\delta^2$, and $p_2 = p_y/\delta$ and neglecting higher order terms in δ, one obtains the following approximate equations of motion:

$$\begin{aligned}
\dot{q}_1 &= -q_1 - q_2, & \dot{q}_2 &= K\delta\, q_2(1 + q_1 + 2p_2), \\
\dot{p}_1 &= p_1 - q_2 p_2, & \dot{p}_2 &= K\delta\left(p_1 - p_2 - p_2^2 - q_1 p_2\right).
\end{aligned} \tag{100}$$

The fixed points become $M_1 = [0, 0, 0, 0]$, $M_2 = [-1, 1, 0, 0]$, and $F = [0, 0, 0, -1]$. For $K\delta \ll 1$ the subsystem (q_1, p_1) is fast, whereas (q_2, p_2) is slow. On the fast time scale (that is, the time scale μ^{-1} in the original, dimensional variables) the fast subsystem approaches the state $q_1 \simeq -q_2$ and $p_1 \simeq q_2 p_2$ which then slowly evolves according to the equations

$$\dot{q}_2 \simeq K\delta\, q_2(1 - q_2 + 2p_2), \quad \dot{p}_2 \simeq K\delta\, p_2(2q_2 - 1 - p_2) \tag{101}$$

that are Hamiltonian, as they follow from the reduced Hamiltonian $H_r(q_2, p_2) = K\delta\, q_2 p_2(1 - q_2 + p_2)$. This universal Hamiltonian appears in the theory of a whole class of *single*-species models in the vicinity of the transcritical bifurcation point [106], see also section 6.2. As $H_r(q_2, p_2)$ is independent of time, it is an integral of motion. The optimal extinction path goes along the zero-energy trajectory $1 - q_2 + p_2 = 0$. Evaluating the action (98) along this trajectory, we find in the leading order: $\mathcal{S}_0 \simeq \left[N\delta^3/(K\delta)\right] \int_1^0 p_2 \mathrm{d}q_2 = N\delta^2/(2K)$. For the MTE of the disease we obtain

$$\ln(\tau)/N \simeq \delta^2/(2K) = [\mu/(2\beta)]\,(1 - \mu_I/\beta)^2\,; \qquad (102)$$

this asymptote is valid when $\mathcal{S}_0 \gg 1$. Many other stochastic multi-population models exhibit a similar time-scale separation close to the bifurcation, defining a universality class [31, 106].

7.1.2. Speeding up endemic disease extinction with a limited vaccination.

A common way of fighting epidemics is vaccination. Mathematical modeling of different aspects of vaccination has attracted considerable attention from mathematicians and, more recently, from physicists [118]. Here we will briefly consider only one aspect of vaccination, following [38]. If there is enough vaccine (and the susceptible individuals are willing to be vaccinated), the infection may be often eradicated 'deterministically'. The amount of available vaccine, however, can be insufficient. The vaccine can be expensive, or dangerous to store in large amounts, or it can be effectively short-lived because of mutations of the infection agent. Finally (and unfortunately), many people nowadays opt out of vaccination programs for reasons unrelated to science. In some cases spontaneous extinction of an endemic disease from a population can still be greatly accelerated even if only a fraction of susceptible individuals are vaccinated [31, 38]. By reducing the number of susceptible individuals, the vaccination perturbs the instanton (which describes the optimal path toward the disease extinction in the absence of vaccination), and leads to a reduction of the 'entropic barrier' to extinction. The mathematical solution of this problem boils down to an optimization problem where one maximizes the reduction of the classical action for given constraints on the vaccine [38]. It turns out that, if the available amount of vaccine is constrained by a given average vaccination rate, the optimal vaccination protocol turns out to be model-independent for a whole class of epidemiological models (including the SI model with population turnover, considered in the previous section). Furthermore, if a vaccination protocol is periodic in time, the optimal protocol represents a periodic sequence of delta-like pulses. The disease extinction rate can strongly depend on the period of this sequence, and display sharp peaks when the vaccination frequency is close to the characteristic frequency of the oscillatory deterministic dynamics of the epidemic in the absence of demographic fluctuations, or to its sub-harmonics [38].

7.1.3. Minimizing the population extinction risk by migration.

In our next example we will follow [47] and evaluate the MTE of a metapopopulation which consists of several local populations occupying separate 'patches'. A local population is prone to extinction due to the demographic noise. A migrating population from another patch can dramatically delay the extinction. What is the optimal migration rate that maximizes the MTE of the whole population? This question was addressed in [47]. The authors considered N local populations of individuals A located on a connected network of patches $i = 1, 2,..., N$ with different carrying capacities. The individuals undergo branching $A \rightarrow 2A$ with rate constant 1 on each patch and annihilation $2A \rightarrow \emptyset$ with rate constant $1/(\kappa_i K)$ on patch i. It is assumed that $K \gg 1$. The parameters $\kappa_i = O(1)$, $i = 1, 2, \ldots, N$, describe the difference among the local carrying capacities $\kappa_i K$. Each individual can also migrate between connected patches i and j with rate constant $\mu_{ij} = \mu_{ji}$. It is assumed that $\mu_{ij} = \mu M_{ij}$, where elements of M_{ij} are of order unity.

The MTE of the meta-population, τ, is exponentially large in K but finite. How does τ depend on the characteristic migration rate μ? At $\mu = 0$ each local population goes extinct separately, and $\tau_{\mu=0}$ is determined by the patch with the greatest carrying capacity, $K_m = K \max_i \{\kappa_i\}$:

$$\ln \tau_{\mu=0}/K \simeq 2(1 - \ln 2)\, \max_i \{\kappa_i\} \qquad (103)$$

see equation (62). At very fast migration, $\mu \rightarrow \infty$, the local populations are fully synchronized: both at the level of the expected local carrying capacities, and at the level of large

fluctuations leading to population extinction. The total carrying capacity of the meta-population, as derived from the deterministic rate equation for this model [47], becomes $\bar{\kappa}K$, where

$$\bar{\kappa} = N^2 / \sum (\kappa_i^{-1}). \tag{104}$$

As a result, at $\mu \to \infty$ the meta-population goes extinct as if it were occupying a single effective patch with the total rescaled carrying capacity $\bar{\kappa}$, that is

$$\ln \tau_{\mu \to \infty}/K \simeq 2(1 - \ln 2)\, \bar{\kappa}. \tag{105}$$

The main result of [47] is that, for unequal κ_i, τ reaches its maximum at a finite (and very small) value of the migration rate. This fact is intimately related to synchronization of the most probable local extinction events that occurs already at very small but positive migration rates. The synchronization makes τ close to that for a single patch with the *combined* carrying capacity $K \sum_i \kappa_i$:

$$\ln \tau_{\mu \to +0}/K \simeq 2(1 - \ln 2) \sum_i \kappa_i. \tag{106}$$

Now let us inspect the MTE in the cases of $\mu = 0, \mu = \infty$ and very small but finite μ, as described by equations (103), (105) and (106), respectively. As $\sum_i \kappa_i \geqslant \max_i\{\kappa_i\}$ and $\sum_i \kappa_i \geqslant \bar{\kappa}$ for any κ_i, the MTE must indeed reach a maximum at a finite value $\mu = \mu_*$, unless all the patches have the same carrying capacity. As was found in [47], $\mu_* \ll 1$ and scales as $1/K$.

Now we expose the theory of [47] in some detail. For simplicity, we will limit ourselves to a system of only two patches. The deterministic rate equations are:

$$\dot{x} = x - x^2 - \mu x + \mu y, \quad \dot{y} = y - \frac{y^2}{\kappa} + \mu x - \mu y, \tag{107}$$

where x and y are the local population sizes rescaled by $\kappa_1 K$, and $\kappa = \kappa_2/\kappa_1$. Equations (107) have two fixed points: the unstable point $x_0 = y_0 = 0$ that describes an empty system, and a stable point $[x_*(\kappa, \mu) > 0, y_*(\kappa, \mu) > 0]$ that describes an established meta-population. At $\mu = 0$ one has $x_* = 1$ and $y_* = \kappa$, whereas for infinitely fast migration, $\mu \to \infty$,

$$x_* = y_* = 2\kappa/(1 + \kappa). \tag{108}$$

The characteristic time t_r of population establishment is determined by the smaller of the two eigenvalues of the linear stability matrix of equations (107) at the fixed point (x_*, y_*).

In the stochastic formulation, the probability $P_{m,n}(t)$ to find m individuals in patch 1 and n individuals in patch 2 evolves in time according to the master equation

$$\begin{aligned}
\dot{P}_{m,n}(t) = \hat{H}P_{m,n} \equiv\ & (m-1)P_{m-1,n} + (n-1)P_{m,n-1} \\
& + \frac{(m+1)(m+2)}{2K}P_{m+2,n} + \frac{(n+1)(n+2)}{2\kappa K}P_{m,n+2} \\
& + \mu(m+1)P_{m+1,n-1} + \mu(n+1)P_{m-1,n+1} \\
& - \left[(1+\mu)(m+n) + \frac{m(m-1)}{2K} + \frac{n(n-1)}{2\kappa K}\right]P_{m,n}.
\end{aligned} \tag{109}$$

The probability $P_{0,0}$ that the meta-population goes extinct by time t obeys the equation

$$\dot{P}_{0,0}(t) = \frac{1}{K}P_{2,0} + \frac{1}{\kappa K}P_{0,2}. \tag{110}$$

At $t \gtrsim t_r$, $P_{m,n}(t)$ is sharply peaked at the local carrying capacities $m_* = Kx_*$ and $n_* = Ky_*$, corresponding to the stable fixed point (x_*, y_*) of the deterministic theory. The subsequent slow decay of $P_{m,n}$ in time is determined by the lowest excited eigenmode $\pi_{m,n}$ of the
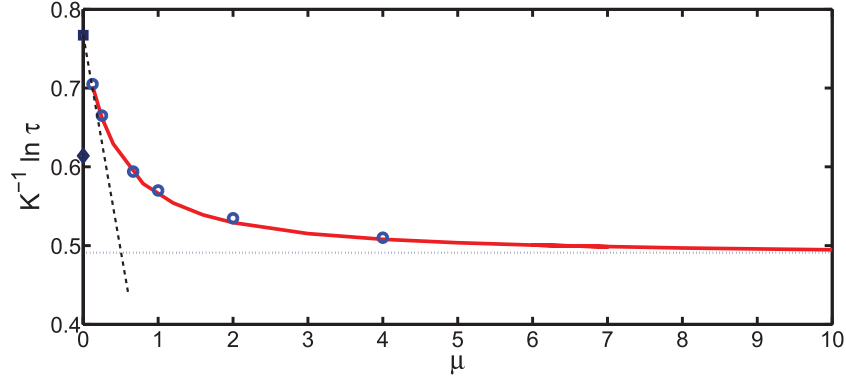
Figure 17. $K^{-1}\ln\tau$ versus the migration rate μ for a two-patch metapopulation and $\kappa = 0.25$. Circles: numerical WKB solutions. Diamond and square: predictions of equations (103) and (105), respectively. Dashed line: prediction of linear theory for $\mu \ll 1$. Dotted line: prediction of the theory for $\mu \gg 1$. The solid line was obtained from a numerical solution of the master equation (109) for $K = 220$. Reprinted figure with permission from [47], copyright 2015 by the American Physical Society.

master equation operator $\hat{H}$: $P_{m,n}(t) \simeq \pi_{m,n}\exp(-t/\tau)$. Simultaneously, a probability peak at $m = n = 0$ grows with time: $P_{0,0}(t) \simeq 1 - \exp(-t/\tau)$. The inverse eigenvalue τ is an accurate approximation to the MTE. Since it is exponentially large with respect to $K \gg 1$, one can neglect the right-hand-side of the eigenvalue problem $\hat{H}\pi_{m,n} = -\pi_{m,n}/\tau$ and consider the quasi-stationary equation $\hat{H}\pi_{m,n} \simeq 0$. Once $\pi_{m,n}$ is found, the MTE can be determined from equation (110):

$$\tau = [\pi_{2,0}/K + \pi_{0,2}/(\kappa K)]^{-1} \tag{111}$$

The WKB ansatz for $\pi_{m,n}$ is

$$\pi_{m,n} = \exp[-KS(x,y)], \tag{112}$$

where $x = m/K$ and $y = n/K$. In the leading order in $1/K \ll 1$ one obtains a zero-energy Hamilton–Jacobi equation $H(x, y, \partial_x S, \partial_y S) = 0$ with classical Hamiltonian

$$
\begin{aligned}
H(x,y,p_x,p_y) = {}& x\left(e^{p_x} - 1\right) + \frac{x^2}{2}\left(e^{-2p_x} - 1\right) + y\left(e^{p_y} - 1\right) + \frac{y^2}{2\kappa}\left(e^{-2p_y} - 1\right) \\
& + \mu x\left(e^{-p_x+p_y} - 1\right) + \mu y\left(e^{p_x-p_y} - 1\right).
\end{aligned}
\tag{113}
$$

The established population corresponds to the fixed point $M = (x_*, y_*, 0, 0)$ of the Hamiltonian flow. Up to a pre-exponent, $\tau \sim \exp(KS)$, where S is the action along the instanton that exits, at time $t = -\infty$, the fixed point M and approaches the fluctuational extinction point F that, for the two-patch branching-annihilation model, is $(0, 0, -\infty, -\infty)$. In general, the instanton and the action along it can be found only numerically. Analytical results are possible in the limits of small and large μ. When $\mu \to +0$ the Hamiltonian (113) becomes separable, and the extinction instanton can be easily found analytically leading to the MTE (106) with $N = 2$. This MTE is exponentially large compared to the one obtained if one neglects migration completely, see equation (103) with $\max\{\kappa_i\} = 1$. The sharp increase of τ once slow migration is allowed results from synchronization of the most probable local extinction paths of the local populations.

The first order correction in μ to the action can be calculated perturbatively, similarly to the case of a weak environmental modulation, see section 6.1.1, by integrating the terms

Table 2. Stochastic predator-prey model. Reprinted table with permission from [44], copyright 2015 by the American Physical Society.

Process	Transition	Rate
Birth of rabbits	$R \rightarrow 2R$	aR
Predation and birth of foxes	$F + R \rightarrow 2F$	$\frac{RF}{\Gamma N}$
Death of foxes	$F \rightarrow \emptyset$	F
Death of rabbits	$R \rightarrow \emptyset$	bR
Competition among rabbits	$2R \rightarrow R$	$\frac{R(R+1)}{2N}$

proportional to μ of the classical Hamiltonian (113) over the unperturbed ($\mu \rightarrow +0$) x- and y-instantons. In the opposite limit, $\mu \rightarrow \infty$, the total population size $Q = x + y$ varies slowly in comparison with the fast migration. The fast variables x and y rapidly adjust to the slow dynamics of Q, staying close to their stationary values for the instantaneous value of Q. Transforming to Q and $q = x$ and associated conjugate momenta as a new set of canonical variables, one arrives at an effective *one-dimensional* Hamiltonian. The action along its instanton can be calculated analytically, and it yields equations (104) and (105) with $N = 2$. It would be interesting to also calculate the subleading correction in $1/\mu$. Figure 17 shows the numerically found S for $\kappa = 0.25$ and different μ. The numerical instantons were found in [47] by using both the shooting method and the iteration method, with very close results.

At $\mu \ll 1$ and $\mu \gg 1$ the numerical results agree with analytical predictions. Figure 17 also compares the WKB results with those of a numerical solution of (a truncated version of) the full master equation (109).

To evaluate the maximum MTE and the optimal migration rate, one needs to resolve the jump of $(\ln \tau)/K$ at $\mu = 0$ predicted by the WKB theory, see equations (103) and (106). The authors of [47] determined the MTE for very small μ by numerically solving the master equation (109) and by performing stochastic simulations. The resulting μ-dependence of the MTE, at $\kappa = 0.25$ and different K, is shown in figure 18. The maximum of τ is observed at a small migration rate μ_* that apparently scales as K^{-1}. This regime is beyond the WKB approximation, and an analytical theory of this regime is currently unavailable.

7.1.4. Multiple routes to extinction. In some population models extinction of a population may occur via more than one scenario. This situation was considered in [44] on the example of a simple predator-prey model, see table 2. This model generalizes the celebrated Lotka–Volterra model [119] by taking into account competition among prey.

By re-interpreting the rabbits (R) as susceptibles (S) and the foxes (F) as infected (I), this model also describes extinction of an endemic disease in an isolated community [44]. As in the SI model with population turnover, a susceptible individual can become infected upon contact with another infected, and the susceptibles and infectives are removed with constant per-capita rates. In the conventional SI model with population renewal, see section 7.1.1, the susceptibles arrive from outside. Instead, in the modified SI model, the susceptibles reproduce by division. Finally, the susceptibles compete for resources, $2S \rightarrow S$, so their population size remains bounded.

Introducing the rescaled population sizes $x = R/N$ and $y = F/N$, one can write the deterministic rate equations as

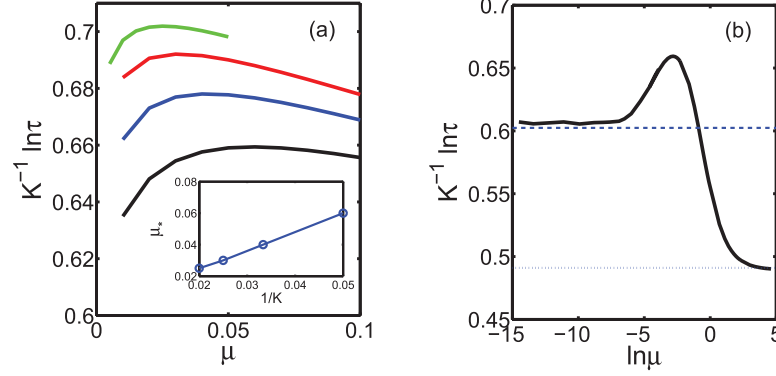$$\dot{x} = (a - b)x - \frac{xy}{\Gamma} - \frac{x^2}{2}, \quad \dot{y} = \frac{xy}{\Gamma} - y. \tag{114}$$

Figure 18. $K^{-1} \ln \tau$ versus μ (a) and $\ln \mu$ (b) for a two-patch meta-population from a numerical solution of the master equation and stochastic simulations. (a) $\kappa = 0.25$ and $K = 20, 30, 40$ and 50 (bottom to top). Inset: the migration rate μ_*, at which the maximum of the MTE is observed, versus K. (b) $\kappa = 0.25$ and $K = 20$; dashed line: equation (103) with $N = 2$, dotted line: equations (104) and (105) with $N = 2$. Reprinted figure with permission from [47], copyright 2015 by the American Physical Society.

These equations are fully characterized by two parameters, $a - b$ and Γ. For $a > b$ and $\Gamma < \Gamma_* \equiv 2(a - b)$, equations (114) have three fixed points corresponding to non-negative population sizes. The fixed point M_1 ($\bar{x}_1 = 0$, $\bar{y}_1 = 0$) describes an empty system. It is a saddle point: attracting in the y-direction (no rabbits), and repelling in the x-direction. The fixed point M_2 [$\bar{x}_2 = \Gamma_*$, $\bar{y}_2 = 0$] describes an established population of rabbits in the absence of foxes. It is also a saddle: attracting in the x-direction (no foxes), and repelling in a direction corresponding to the introduction of a small number of foxes into the system. The third fixed point M_3 [$\bar{x}_3 = \Gamma$, $\bar{y}_3 = \Gamma(\Gamma_* - \Gamma)/2$] is attracting and describes the 'coexistence' state. It is either a stable node (for $\Gamma > \Gamma_0 = 4(\sqrt{1 + a - b} - 1)$), or a stable focus (for $\Gamma < \Gamma_0$). Note that $\bar{y}_3$ is a non-monotonic function of Γ. It vanishes at $\Gamma = 0$ and $\Gamma = \Gamma_*$, and reaches a maximum, $\Gamma_*^2/8$, at $\Gamma = \Gamma_*/2$ corresponding to the optimal predation rate. Figure 19 shows two examples of deterministic trajectories: for $\Gamma_0 < \Gamma < \Gamma_*$ (a) and for $\Gamma < \Gamma_0$ (b).

Reintroducing the stochasticity, both populations ultimately go extinct, but now there are two distinct extinction routes. In the first, *sequential* route the predators (or infectives) go extinct first, whereas the prey (or susceptibles) typically thrive for a long time and only then go extinct. In the second, *parallel* route the prey (or susceptibles) go extinct first, causing a rapid extinction of the predators (or infected).

Let $P_{m,n}(t)$ be the probability to observe, at time t, m rabbits and n foxes, where $m, n = 0, 1, 2 \ldots$. The evolution of $P_{m,n}(t)$ is described by the master equation

$$
\begin{aligned}
\dot{P}_{m,n} = \hat{H} P_{m,n} &= a[(m - 1)P_{m-1,n} - m P_{m,n}] \\
&+ (1/\Gamma N)[(m + 1)(n - 1)P_{m+1,n-1} - mn P_{m,n}] \\
&+ b[(m + 1)P_{m+1,n} - m P_{m,n}] + (n + 1)P_{m,n+1} - n P_{m,n} \\
&+ (1/2N)[(m + 1)m P_{m+1,n} - m(m - 1)P_{m,n}],
\end{aligned}
\tag{115}
$$

where $P_{i,j} = 0$ when any of the indices is negative. Before delving into the WKB analysis of this master equation, let us discuss the probability flow in this system. At large N and at sufficiently long times, the probability distribution $P_{m,0}(t)$ with $m > 0$, which describes the dynamics of rabbits conditional on prior extinction of the foxes, is a one-dimensional distribution peaked at the fixed point M_2 of the deterministic theory. The probability distribution $P_{m,n}(t)$
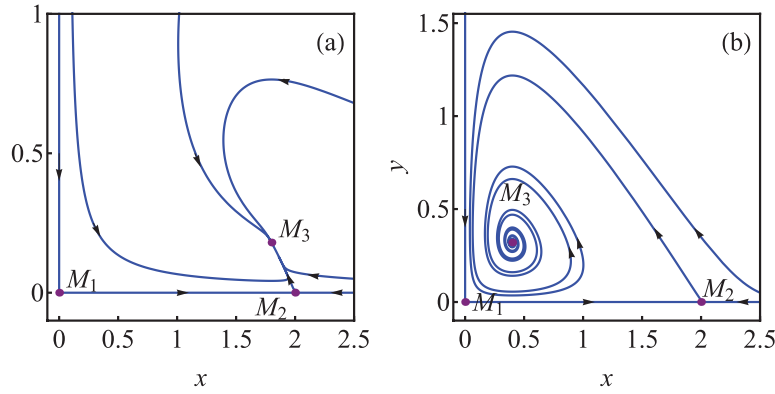
Figure 19. Deterministic trajectories of the predator-prey model (see table 2) for $a = 2$, $b = 1$ and two different values of Γ. (a) $\Gamma = 1.8$: M_3 is a stable node. (b) $\Gamma = 0.4$: M_3 is a stable focus. Reprinted figure with permission from [44], copyright 2015 by the American Physical Society.

with $m, n > 0$ (which describes the long-time dynamics of the coexisting rabbits and foxes) is a two-dimensional distribution peaked at the fixed point M_3. Finally, the extinction probability of both sub-populations $P_{0,0}(t)$ corresponds to a Kronecker-delta probability distribution. Not only the structure, but the long-time dynamics of these three distributions are different. To emphasize this point, the authors of [44] defined the total 'probability contents' of the vicinities of each of the fixed points M_1, M_2 and M_3:

$$\mathcal{P}_1(t) = P_{0,0}(t), \tag{116}$$

$$\mathcal{P}_2(t) = \sum_{m=1}^{\infty} P_{m,0}(t), \tag{117}$$

$$\mathcal{P}_3(t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} P_{m,n}(t). \tag{118}$$

At $N \gg 1$ and $t \gg t_r$ the sums in equations (117) and (118) are mostly contributed to by close vicinities of the points M_2 and M_3, respectively. The long-time evolution of $\mathcal{P}_1(t)$, $\mathcal{P}_2(t)$ and $\mathcal{P}_3(t)$ is described by the effective *three-state* master equation:

$$\dot{\mathcal{P}}_1 = r_{21}\mathcal{P}_2 + r_{31}\mathcal{P}_3, \quad \dot{\mathcal{P}}_2 = -r_{21}\mathcal{P}_2 + r_{32}\mathcal{P}_3, \quad \dot{\mathcal{P}}_3 = -(r_{31} + r_{32})\mathcal{P}_3, \tag{119}$$

where r_{ij} is the (yet unknown) transition rate from the vicinity of the fixed point i to the vicinity of the fixed point j. Let the initial condition correspond to the coexistence state around M_3:

$$[\mathcal{P}_1(0), \mathcal{P}_2(0), \mathcal{P}_3(0)] = (0, 0, 1). \tag{120}$$

Then the solution of equation (119) is

$$\mathcal{P}_1(t) = 1 + \frac{r_{32}\,\mathrm{e}^{-r_{21}t} + (r_{31} - r_{21})\,\mathrm{e}^{-(r_{31}+r_{32})t}}{r_{21} - r_{31} - r_{32}}, \tag{121}$$

$$\mathcal{P}_2(t) = \frac{r_{32}\left[\mathrm{e}^{-(r_{31}+r_{32})t} - \mathrm{e}^{-r_{21}t}\right]}{r_{21} - r_{31} - r_{32}}, \tag{122}$$

$$\mathcal{P}_3(t) = e^{-(r_{31}+r_{32})t}. \tag{123}$$

Once the transition rates r_{31}, r_{32} and r_{21} are known, equations (121)–(123) provide a useful 'coarse-grained' description of this system in terms of the long-time evolution of the probabilities to observe the coexistence state around M_3, the fox-free state around M_2 and the extinction state at M_1. In particular, the MTE of foxes is $\tau_F \simeq (r_{31} + r_{32})^{-1}$, whereas the MTE of both populations is

$$\tau_{RF} = \int_0^\infty \mathrm{d}t\, t\, \dot{\mathcal{P}}_1(t) = \frac{r_{21} + r_{32}}{r_{21}(r_{31} + r_{32})}. \tag{124}$$

The transition rate r_{21} comes from solving the one-population problem $R \rightleftarrows 2R$ and $R \to \emptyset$; it is known with a high accuracy [30, 35]. To evaluate the transition rates r_{31} and r_{32}, the authors of [44] used the real-space WKB theory. The classical Hamiltonian takes the form

$$H(x,y,p_x,p_y) = ax\left(e^{p_x}-1\right) + bx\left(e^{-p_x}-1\right) + \frac{xy}{\Gamma}\left(e^{p_y-p_x}-1\right) + y\left(e^{-p_y}-1\right) + \frac{x^2}{2}\left(e^{-p_x}-1\right). \tag{125}$$

The Hamiltonian flow, generated by this Hamiltonian, has five zero-energy fixed points:

$$M_1 = (0,0,0,0); \qquad M_2 = (\Gamma_*,0,0,0); \qquad M_3 = [\Gamma, \Gamma(\Gamma_* - \Gamma)/2, 0, 0];$$
$$F_1 = [0,0,\ln(b/a),0]; \quad F_2 = [\Gamma_*, 0, 0, -\ln(\Gamma_*/\Gamma)]. \tag{126}$$

The zero-momentum fixed points M_1, M_2 and M_3 correspond to the three fixed points of the deterministic equations (114). The two other fixed points, F_1 and F_2, are fluctuational fixed points describing a fox-free state at a non-zero number of rabbits (F_2) and an empty system (F_1). The two different routes to extinction—the sequential and the parallel—are encoded in two different instantons. Both start from the fixed point M_3 which describes the established populations, but end in a different fluctuational fixed point, F_1 or F_2. The instantons and the action along each of the instantons can be found numerically [44]. Figure 20 shows typical examples of the numerically found instantons M_3F_1 and M_3F_2 in the case when M_3 is a node. Figure 21 refers to the case when M_3 is a focus. The actions S_{31} and S_{32} along these (zero-energy) instantons,

$$S_{31} = \int_{M_3}^{F_1} p_x \mathrm{d}x + p_y \mathrm{d}y \quad \text{and} \quad S_{32} = \int_{M_3}^{F_2} p_x \mathrm{d}x + p_y \mathrm{d}y,$$

yield, with exponential accuracy, the transitions rates r_{31} and r_{32}:

$$r_{31} \sim \exp(-NS_{31}), \quad r_{32} \sim \exp(-NS_{32}). \tag{127}$$

As the actions along the different instantons are different, and $N \gg 1$, there is usually a great (exponential) disparity between the transition rates: $r_{21} \ll r_{31} \ll r_{32}$, implying that the sequential extinction route (foxes first, rabbits second) is usually much more likely than the parallel extinction route. In other words (and somewhat surprisingly), the predators are usually more prone to extinction than their prey.

7.1.5. Epidemic fadeout. An infectious disease can disappear from a population immediately after the first infection outbreak [2, 113]. Such an 'epidemic fade-out' happens if the epidemic dynamics are oscillatory at the level of the deterministic theory, and the number of infected individuals at the end of the first outbreak of the disease is relatively small so that fluctuations in the disease transmission and in the recovery or removal of the infected can 'switch off' the disease. Epidemic
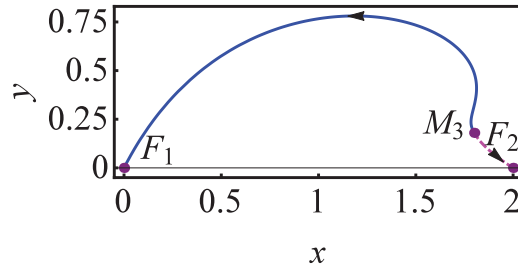
Figure 20. Instantons M_3F_1 and M_3F_2 for $a = 2$, $b = 1$ and $\Gamma = 1.8$, when the fixed point M_3 is a stable node. Shown are the x, y-projections of the four-dimensional instantons. Reprinted figure with permission from [44], copyright 2015 by the American Physical Society.

fade-out has been mostly addressed by epidemiologists via stochastic simulations. (One exception is [113] which reported important early analytical results.) To determine the probability of the epidemic fadeout, [73] considered the SI model with population turnover, see table 1, therefore the master equation coincides with equation (93). The initial condition, $P_{n,m}(t = 0) = \delta_{n,N}\delta_{m,m_0}$, used in [73], describes introduction of m_0 infected individual into a steady-state susceptible population. One boundary condition reflects the fact that $m = 0$ is, for any n, an absorbing state. Being interested in epidemic fade-out, one should exclude from consideration all stochastic trajectories that do not reach the extinction boundary $m = 0$ immediately after the first outbreak. This can be achieved by introducing an artificial absorbing boundary [58, 73, 113].

In contrast to the endemic fadeout, which is studied assuming quasi-stationarity of $P_{n,m}$, the epidemic fadeout occurs on a fast time scale determined by the deterministic equations. Still, a stationary formulation can be obtained if one uses the master equation (93) to derive an exact stationary equation for the *mean residence time* of the population in a certain state (n, m), where $m > 0$: $T_{n,m} = \int_0^\infty P_{n,m}(t)\, \mathrm{d}t$ [73]. (Alternatively, one could write a recursive equation for the extinction probability starting from the state n, m, as was done in section 3.) The accumulated extinction probability $\mathcal{P}_n$ from the state $(n, 1)$ becomes $\mathcal{P}_n = \mu_I T_{n,1}$, and the total extinction probability is $\mathcal{P} = \sum_n \mathcal{P}_n$. Integrating equation (93) over t from 0 to ∞ and using the equality $P_{n,m}(t = \infty) = 0$ and the initial condition, one obtains for $T_{n,m>0}$:

$$\sum_{n',m'>0} M_{n,m;\,n',m'}\, T_{n',m'} + \delta_{n,n_0}\, \delta_{m,m_0} = 0, \tag{128}$$

where $M_{n,m;n',m'}$ was defined in equation (93). For $N \gg 1$ this stationary equation can be approximately solved by the WKB ansatz $T_{n,m} = a(x,y)\, \mathrm{e}^{-NS(x,y)}$, where a and S are smooth functions of $x = n/N - 1$ and $y = m/N$ [73]. Remarkably, the ensuing classical Hamiltonian is the same as in equation (95) up to the canonical transformation from the momentum space to the real space.

If only a few infected individuals are introduced into an infection-free population then one has to find an instanton that exits, at $t = -\infty$, the deterministic fixed point corresponding to the infection-free steady state $[S = N, I = 0, p_S = 0, p_I = 0]$ and enters, at $t = \infty$, the fluctuational fixed point for which $S = N$, $I = 0$ and $p_S = 0$, but $p_I \neq 0$ [73]. As was found in [73], such an instanton exists if and only if the endemic fixed point, predicted by the rate equations, is a focus rather than a node. In fact, there exist *multiple* heteroclinic trajectories between the same two points in this case. They can be classified by whether their x, y-projections exhibit a single loop, two loops, three loops, etc [73]. A single-loop instanton corresponds to a disease fade-out immediately after the first outbreak. A two-loop instanton corresponds to a fade-out
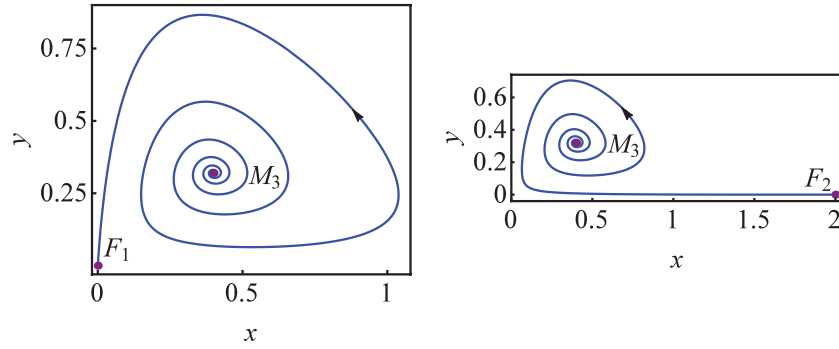
Figure 21. (Left) instanton M_3F_1 and (right) instanton M_3F_2, for $a = 2$, $b = 1$ and $\Gamma = 0.4$, when the fixed point M_3 is a stable focus. Shown is the x, y-projection of the four-dimensional instantons. Reprinted figure with permission from [44], copyright 2015 by the American Physical Society.

immediately after the second outbreak, etc. An example of the x, y-projection of the one-loop instanton is shown in figure 22, where $x = S/N$ and $y = I/N$. In the leading WKB order one has $\mathcal{P} \sim \mathcal{P}_{n=N}$, therefore one again needs to calculate the accumulated action $\mathcal{S}_0$ along the instanton. The numerics can be done by shooting. It is convenient to introduce the same rescaled parameters $K = \beta/\mu$ and $\delta = 1 - \mu_I/\beta$ as in section 7.1.1. As shown in [73], the regime of $K\delta \gg 1$ can be investigated analytically. In this regime the epidemic fadeout instanton closely follows the deterministic trajectory, and departs from it only when $y = I/N$ becomes much smaller than unity, see figure 22. Further, for $K\delta \gg 1$ the fluctuations of the number of susceptibles can be neglected, and one can Taylor-expand the Hamiltonian in $p_x \ll 1$ and truncate the expansion at first order. Under these conditions the main contribution to the action comes from a narrow region of x where the instanton quickly departs from the deterministic orbit. The resulting action looks especially simple if, in addition to the inequality $K\delta \gg 1$, one also demands $\delta \ll 1$. Here one obtains $-\ln \mathcal{P}/N \simeq \mathcal{S}_0 \simeq (2\delta^5/\pi e^4 K)^{1/2}(e/2)^{-K\delta}$. We refer the reader to [73] for details. We also note that, slightly above the node-focus transition of the deterministic theory, the epidemic fadeout instanton first closely follows the deterministic trajectory, and then rapidly approaches the *endemic* fadeout instanton considered in section 7.1.1. The intimate connection between the three heteroclinic trajectories in the four-dimensional phase space is fascinating. It deserves a further study and is likely to be universal for a whole class of models.

7.1.6. Extinction: two more examples. It was realized in [34] that the population extinction rate in multi-population systems may exhibit *fragility*, where the WKB action (that determines the effective barrier to extinction) depends on some of the elemental transition rates non-analytically. As an example, the authors of [34] considered the classic SIS (susceptible-infectious-susceptible) model and its extension accounting for population turnover. The classic SIS model includes the following basic reactions [6, 8, 10]. Upon interaction with an infected individual, a susceptible can be infected, while an infected individual can recover:

$$S + I \xrightarrow{\beta} I + I, \quad I \xrightarrow{\gamma} S. \tag{129}$$

The SIS model with population turnover includes, in addition to these transitions, the renewal of susceptibles, $\emptyset \xrightarrow{\mu} S$. The authors of [34] observed that the WKB action, corresponding to extinction of the endemic disease from the population, experiences a jump when the population renewal is allowed, even in the limit of $\mu \to 0$. To better understand the extinction rate
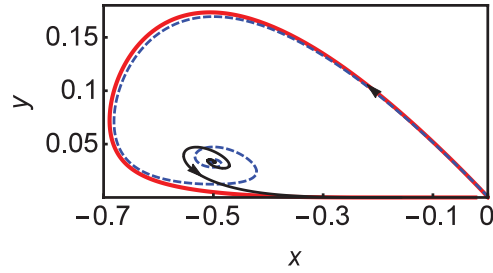
Figure 22. An epidemic outbreak on the *xy*-plane as predicted by the deterministic rate equations for the SI model (dashed line) and the epidemic fade-out instanton (thick solid line). For comparison, the thin solid line shows the *endemic* fade-out instanton considered in section 7.1.1. The rescaled parameters are $K = \beta/\mu = 30$ and $\delta = 1 - \mu_I/\beta = 0.5$. For $K\delta \gg 1$ the epidemic fade-out instanton closely follows the deterministic trajectory and departs from it at $y \ll 1$. Reprinted figure with permission from [73], copyright 2015 by the American Physical Society.

fragility, the subsequent work [120] introduced the concept of time-resolved extinction rate. When μ is small, there is a time-scale separation in the system which enables one to define a short-time quasistationary extinction rate W_1 and a long-time quasistationary extinction rate W_2, and to follow the transition between W_1 and W_2 as it develops in time. Importantly, W_1 and W_2 coincide with the extinction rates when the population turnover is absent and present but very slow, respectively. The extinction rate fragility, discovered in [34], manifests itself in the exponentially large disparity between W_1 and W_2.

An interesting variant of the problem of extinction of a stochastic metapopulation was considered in [121]. The authors considered $N \gg 1$ patches supporting a local birth-death-competition population dynamics. In addition, an individual of each patch can migrate into a pool of migrants, and the reverse transition—from the pool to the patch—also occurs. The authors studied a long-lived quasi-stationary state of the system and employed the real-space WKB method to evaluate the MTE of the whole population. They also found the conditions under which their individual-based model reduces to several metapopulation models, previously used *ad hoc* by ecologists.

7.1.7. Switching in a two-population system. The calculations of the MST in the model of a self-regulating gene, presented in section 2.2.2, can be generalized by explicitly accounting for mRNA dynamics, and this generalization may dramatically affect the MST [43, 91], see below. Here we follow [51], where this generalization was made on an example of a one-state DNA model with positive feedback (see also [91]). The fact that the DNA has only one state means that the mRNA molecules are being constantly transcribed, as opposed to multi-state models [43], see figure 23(a). Regardless of the number of DNA states, we assume that the protein of interest positively regulates the transcription of the mRNA molecules that are responsible for the protein's production, giving rise to a positive feedback loop and a genetic switch.

The deterministic rate equations for the average copy numbers of mRNA, *m*, and proteins, *n*, are [51]:

$$\dot{m} = f(n) - \gamma m,$$
$$\dot{n} = b\gamma m - n. \tag{130}$$

Here transcription of mRNA occurs at a rate $f(n)$ which depends on the current protein number, γ is the mRNA degradation rate, $b\gamma$ is the protein translation rate, and all rates are rescaled by the protein's degradation rate or cell division rate [96]. Furthermore, in the limit of short-lived

mRNA, $\gamma \gg 1$, b is the mean number of proteins made per mRNA [91], see below. We assume that the nonlinear function $f(n)$ introduces a positive feedback leading to three deterministic fixed points: $n_1 \equiv n_{\mathrm{off}} < n_2 < n_3 \equiv n_{\mathrm{on}}$, where n_{off} and n_{on} are attracting fixed points corresponding to the off and on phenotypic states, see section 2.2.2, while n_2 is a repelling fixed point. Let us also denote by $N \equiv n_{\mathrm{on}} \gg 1$ the protein abundance in the on state, see below.

Being interested in the MST between, say, the on and off states, we turn to the master equation for $P_{m,n}(t)$, the probability to find m mRNA molecules and n proteins:

$$\dot{P}_{m,n} = \hat{H}P_{m,n} = f(n)[P_{m-1,n} - P_{m,n}] + \gamma[(m+1)P_{m+1,n} - mP_{m,n}]$$
$$+ b\gamma m(P_{m,n-1} - P_{m,n}) + (n+1)P_{m,n+1} - nP_{m,n}. \quad (131)$$

To evaluate the MST, we set $P_{m,n} = \pi_{m,n} \mathrm{e}^{-t/\tau}$ and employ the WKB ansatz for the QSD, $\pi_{m,n} = \pi(x,y) \sim \mathrm{e}^{-NS(x,y)}$, where $x = m/N$ and $y = n/N$. In the leading WKB order the problem reduces to finding the proper instanton of the Hamiltonian [51]

$$H = y(\mathrm{e}^{-p_y} - 1) + b\gamma x(\mathrm{e}^{p_y} - 1) + \gamma x(\mathrm{e}^{-p_x} - 1) + \tilde{f}(y)(\mathrm{e}^{p_x} - 1), \quad (132)$$

where $\tilde{f}(y) = f(n)/N$. The instanton can be found numerically, e.g. via the shooting method [28]. Then one computes the action $S(x,y) = \int p_x \mathrm{d}x + p_y \mathrm{d}y = \int^t (p_x\dot{x} + p_y\dot{y})\mathrm{d}t$ along the optimal path, and evaluates the MST.

One regime where analytical progress is possible is that of short-lived mRNA, $\gamma \gg 1$, when there is time-scale separation between the mRNA and proteins [91, 96]. Here $x(t)$ and $p_x(t)$ equilibrate rapidly, and one can express x and p_x via the protein concentration and momentum (see also [110]). In this way one finds [51] $\mathrm{e}^{-p_x} = b(1 - \mathrm{e}^{p_y}) + 1$, and $x = [\tilde{f}(y)/\gamma][b(1 - \mathrm{e}^{p_y}) + 1]^{-2}$, which gives rise to a *reduced* protein-only Hamiltonian [51]

$$H_{\mathrm{red}}(y, p_y) = y(\mathrm{e}^{-p_y} - 1) + \tilde{f}(y)\frac{b(\mathrm{e}^{p_y} - 1)}{b(1 - \mathrm{e}^{p_y}) + 1}. \quad (133)$$

At this point we pause and examine the effect of explicitly incorporating the mRNA species into the gene-expression dynamics. Despite adiabatic elimination of the mRNA species, the reduced Hamiltonian (133) differs from the 1D Hamiltonian (43). In the former, mRNA noise is manifested through the nontrivial dependence of the protein production term on the momentum [43, 51], which effectively accounts for the fact that the proteins are produced in geometrically distributed bursts with mean b [43, 69, 97], see figure 24. As expected, the Hamiltonians coincide in the limit of $b \to 0$ (here $b\tilde{f}$ is equivalent to $\tilde{f}$ in section 2.2.2), which corresponds to a constant burst size [43, 69].

Going back to our reduced Hamiltonian, the instanton is [51]

$$p_y(y) = \ln \frac{(b+1)y}{b[y + \tilde{f}(y)]}. \quad (134)$$

Calculating the reduced action $S(y) = \int^y p_y(y')\mathrm{d}y'$, which can be done explicitly for any function $\tilde{f}(y)$, gives rise to the protein-only QSD, $\pi(y) \sim \mathrm{e}^{-NS(y)}$. (Note that here while calculating the action S we have neglected the contribution from $\int p_x \mathrm{d}x$, which scales as $\mathcal{O}(\gamma^{-1})$ compared to $\int p_y \mathrm{d}y$ [43].) As a result, the MST, say from the on state to the off state, is [51]

$$\tau_{\mathrm{on}\to\mathrm{off}} \sim \mathrm{e}^{N[S(y_2) - S(y_{\mathrm{on}})]}. \quad (135)$$

The calculation of $\tau_{\mathrm{off}\to\mathrm{on}}$ is done in a similar way. Here the accumulated action, $S(y_2) - S(y_{\mathrm{off/on}})$, turns out to be smaller than the accumulated action obtained in the protein-only model for any $b > 0$, see section 2.2.2. As a result, an account for mRNA noise exponentially decreases the MST compared to the protein-only case, see figure 24.

The real-space WKB method can be also used in more complicated scenarios of gene regulatory networks in the presence of multiple DNA states, a mechanism which is used to control mRNA production in a more efficient manner. Examples of such analysis were presented in [43] for a two-state system with nonlinear feedback, and in [45, 49] for multiple DNA-state systems.

Let us briefly review, following [43], the case of a two-state positive feedback switch, which was experimentally shown to describe biological switching [122]. We consider a two-state gene-expression model, where transitions between a transcriptionally active and inactive promoter are controlled by the protein copy number n via positive feedback, see figure 23(a). The transition rates into the active and inactive states, respectively, can be chosen to be

$$
k_{\text{on}}(n) = k_0^{\min} + (k_0^{\max} - k_0^{\min})\frac{n^{h_1}}{n^{h_1} + n_{50}^{h_1}}, \quad k_{\text{off}}(n) = k_1^{\min} + (k_1^{\max} - k_1^{\min})\frac{n^{h_2}}{n^{h_2} + n_{50}^{h_2}}.
$$

(136)

While the analytical approach holds for generic functions $k_{\text{on}}(n)$ and $k_{\text{off}}(n)$, the rates (136) were shown to be biologically relevant, e.g. in the *lac* operon [89]. As before, we denote the stable fixed points by N_{off} and N_{on}, while the unstable point is N_0.

In the case of the two-state model, the master equation has to be written for $P_{m,n}$ and $Q_{m,n}$ representing the probability distribution functions of having m mRNAs and n proteins at time t with the promoter in the inactive and active state, respectively, see [43] for details. Using the WKB machinery, and assuming short-lived mRNA, $\gamma \gg 1$, one can adiabatically eliminate the mRNA species and arrive at a reduced two-state Hamiltonian $H_r^{(2)}(y, p_y)$ [43]:

$$
H_r^{(2)} = (e^{-p_y} - 1)\left\{ y + \left[y + \frac{e^{p_y}}{b(e^{p_y} - 1) - 1}\right]\left[\frac{\tilde{k}_{\text{on}}(y) - y(e^{-p_y} - 1)}{\tilde{k}_{\text{off}}(y)}\right]\right\},
$$

(137)

where $\tilde{k}_{\text{on/off}}(y) = k_{\text{on/off}(n)}/K$, and $K = ab$ is the typical number of proteins in the on state. One finds the instanton $p_y(y)$, see figure 23(b), from the equation $H = 0$, and calculates $S(y) = \int^y p_y(y')\mathrm{d}y'$. Setting $\tilde{k}_{\text{off}}(y) = 0$, which corresponds to a one-state positive feedback switch, and by properly rescaling $\tilde{k}_{\text{on}}(y)$, one recovers equation (134). Finally, having calculated the action $S(y)$, the MST from the off→on states (and similarly from the on→off) is given by

$$
\ln \tau_{\text{off}\to\text{on}} \simeq K\Delta S_{\text{off}},
$$

(138)

where $\Delta S_{\text{off}} = S(y_0) - S(y_{\text{off}})$, $y_{\text{off}} = N_{\text{off}}/K$ and $y_0 = N_0/K$. Figure 24 compares the analytical predictions of [43] with Monte Carlo simulations (see figure 23(c)), and an excellent agreement is observed. Figure 24 also demonstrates that mRNA noise can exponentially decrease the MST between different phenotypic states, see also [91].

7.2. Established populations reside in a vicinity of a limit cycle

Some populations exhibit persistent oscillations in their sizes [119, 123–127]. At the level of the deterministic theory, these oscillations can often by described by a stable limit cycle in the space of population sizes. One deterministic model that shows this feature—an extension of the celebrated Lotka–Volterra model [119]—is due to Rosenzweig and MacArthur [128]. Qualitatively similar models are used in epidemiology—for a description of the oscillatory dynamics of susceptible and infected populations during an epidemic [2, 8, 129] and the oscillatory dynamics of tumor growth [130, 131]. Similar models describe oscillatory
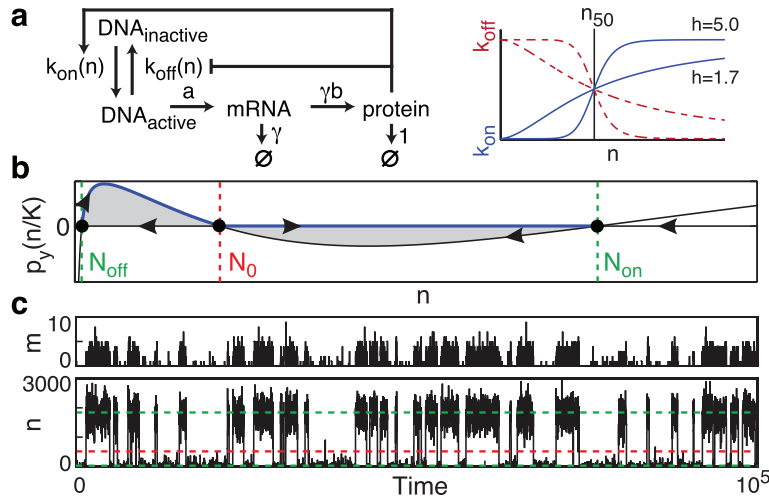
Figure 23. (a) A model for a two-state positive feedback network. Transcription of mRNA and translation of proteins are modeled as first-order processes with rates a and $b\gamma$, respectively. The mRNA and protein molecules also undergo first-order degradation with rates γ and 1 (all the rates are rescaled by the protein decay rate). The feedback functions $k_{on}(n)$ and $k_{off}(n)$ control promoter transitions. (b) The momentum p_y, obtained from solving $H_r^{(2)}(y, p_y) = 0$ (where H is given by equation (137)) versus protein copy number n. The thick line indicates the off→on instanton, the thin line indicates the on→off instanton, while the shaded areas correspond to the switching actions. (c) Time-dependent fluctuations in mRNA (denoted by m), and protein (denoted by n) copy-numbers, in a typical Monte Carlo trajectory undergoing switching. In (b) and (c) $K = ab = 2400$, $b = 22.5$, $h_1 = h_2 = 2$, $n_{50} = 1000$, $k_0^{min} = k_1^{min} = a/100$, $k_0^{max} = k_1^{max} = a$, and $\gamma = 50$. Reprinted figure with permission from [43], copyright 2015 by the American Physical Society.

chemical reactions [59, 132]. Population extinction, driven by demographic noise, acquires new features in this case, as has been shown in [57] on the example of a stochastic version of the Rosenzweig–MacArthur (RMA) model, see table 3. The model differs from that used in [44] (see table 2) in two aspects. First, in the absence of the predators, the prey is immortal. Second, for a very large prey population, the predation rate saturates so as to describe satiation of the predators.

The deterministic equations for the RMA model can be written as

$$\dot{x} = ax - \frac{x^2}{2} - \frac{\sigma xy}{1 + \sigma\tau x}, \quad \dot{y} = -y + \frac{\sigma xy}{1 + \sigma\tau x}, \tag{139}$$

where $x = R/N$, $y = F/N$, and $\sigma = sN = \mathcal{O}(1)$. These equations have been extensively studied [128, 133, 134]. When

$$0 < \tau < 1, \quad \sigma > \sigma_0 = \frac{1}{2a(1-\tau)}, \tag{140}$$

equations (139) have three fixed points describing nonnegative population sizes. The fixed point M_1 ($\bar{x}_1 = 0, \bar{y}_1 = 0$) describes an empty system. It is a saddle point: attracting in the y-direction (when there are no rabbits in the system), and repelling in the x-direction. The fixed point M_2 ($\bar{x}_2 = 2a, \bar{y}_1 = 0$) describes a steady-state population of rabbits in the absence
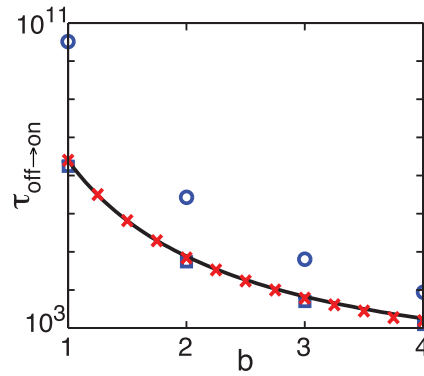
Figure 24. (a) The MST $\tau_{\text{off}\rightarrow\text{on}}$ as a function of the mean burst size b for the two-state model presented in figure 23. Here $K = ab = 2400$, $h_1 = h_2 = 2$, $n_{50} = 720$, $k_0^{\min} = k_1^{\min} = a/100$, $k_0^{\max} = k_1^{\max} = a$ and $\gamma = 50$. Shown are results of Monte-Carlo simulations, see figure 23(c), of the full two-component model (x's), simulations of the protein-only model with geometrically-distributed bursts with mean b ($\square$) and simulations of the protein-only model with constant burst size b ($\circ$). The line is the theoretical prediction. It is evident that, while incorporating the protein production via geometrically-distributed bursts effectively accounts for the mRNA noise, ignoring the mRNA noise exponentially increases the MST. Reprinted figure with permission from [43], copyright 2015 by the American Physical Society.

of foxes. It is also a saddle: attracting in the x-direction (when there are no foxes), and repelling in a direction corresponding to the introduction of a few foxes into the system. The third fixed point M_3 is given by $(\bar{x}_3, \bar{y}_3)$, where

$$\bar{x}_3 = \frac{1}{\sigma(1-\tau)}, \quad \bar{y}_3 = \frac{2a\sigma(1-\tau)-1}{2\sigma^2(1-\tau)^2}, \tag{141}$$

and describes a non-oscillatory coexistence of the rabbits and foxes. For

$$\sigma_0 < \sigma < \bar{\sigma} = \frac{\frac{a\tau(1+\tau)}{2(1-\tau)} - 1 - \sqrt{1 + a\frac{1+\tau}{1-\tau}}}{a^2\tau^2 - 4a(1-\tau)}, \tag{142}$$

M_3 is a stable node. For

$$\bar{\sigma} < \sigma < \sigma^* = \frac{1+\tau}{2a\tau(1-\tau)} \tag{143}$$

it is a stable focus. Finally, for $\sigma > \sigma^*$, M_3 is unstable, and a stable limit cycle appears around it. A Hopf bifurcation occurs at $\sigma = \sigma^*$. Figure 25 shows the behavior of the deterministic model for $\bar{\sigma} < \sigma < \sigma^*$ (a) and for $\sigma > \sigma^*$ (b).

Reintroducing the noise, one can see that, following the population establishment (either around the fixed point, or around the limit cycle), there are two extinction routes in this model. In the first (parallel) route all the rabbits are eaten by the foxes, followed by a quick extinction of the foxes. In the second route the foxes go extinct, while the (immortal) rabbits reach a steady state. Smith and Meerson [57] considered these two routes for $\sigma > \sigma^*$, when the

Table 3. Stochastic Rosenzweig–MacArthur model. Reprinted figure with permission from [57], copyright 2015 by the American Physical Society.

Process	Transition	Rate
Birth of rabbits	$R \to 2R$	aR
Predation and birth of foxes	$F + R \to 2F$	$\frac{sRF}{1+s\tau R}$
Death of foxes	$F \to \emptyset$	F
Competition among rabbits	$2R \to R$	$\frac{R(R+1)}{2N}$

established populations are distributed in the vicinity of the stable limit cycle of the deterministic theory.

Again, we denote by $P_{m,n}(t)$ the probability to observe m rabbits and n foxes at time t. The master equation is

$$\dot{P}_{m,n} = \hat{H} P_{m,n} = a\left[(m-1)P_{m-1,n} - mP_{m,n}\right]$$
$$+ \frac{\sigma(m+1)(n-1)}{N+\sigma\tau(m+1)} P_{m+1,n-1} - \frac{\sigma mn}{N+\sigma\tau m} P_{m,n}$$
$$+ (n+1)P_{m,n+1} - nP_{m,n} + (1/2N)[(m+1)mP_{m+1,n} - m(m-1)P_{m,n}], \tag{144}$$

where $P_{m,n} = 0$ when any of the indices is negative. As in section 7.1.4, one can define the 'effective probability contents' of the vicinities of the fixed points M_1 and M_2 via equations (116) and (117). In its turn, equation (118) now defines the effective probability content of the limit cycle. The long-times dynamics of $\mathcal{P}_1(t)$, $\mathcal{P}_2(t)$ and $\mathcal{P}_3(t)$ are given by the three-state master equation [57]

$$\dot{\mathcal{P}}_1 = \mathcal{R}_1 \mathcal{P}_3, \quad \dot{\mathcal{P}}_2 = \mathcal{R}_2 \mathcal{P}_3, \quad \dot{\mathcal{P}}_3 = -(\mathcal{R}_1 + \mathcal{R}_2)\mathcal{P}_3, \tag{145}$$

where $\mathcal{R}_1$ and $\mathcal{R}_2$ are the extinction rates along the first and second extinction route, respectively. For the initial condition $[\mathcal{P}_1(0), \mathcal{P}_2(0), \mathcal{P}_3(0)] = (0,0,1)$ the solution of equations (145) is

$$\mathcal{P}_1(t) = \frac{\mathcal{R}_1 \left[1 - e^{-(\mathcal{R}_1+\mathcal{R}_2)t}\right]}{\mathcal{R}_1 + \mathcal{R}_2}, \tag{146}$$

$$\mathcal{P}_2(t) = \frac{\mathcal{R}_2 \left[1 - e^{-(\mathcal{R}_1+\mathcal{R}_2)t}\right]}{\mathcal{R}_1 + \mathcal{R}_2}, \tag{147}$$

$$\mathcal{P}_3(t) = e^{-(\mathcal{R}_1+\mathcal{R}_2)t}. \tag{148}$$

As one can see from equations (146)–(148), the long-time behavior of the system is determined by the extinction rates $\mathcal{R}_1$ and $\mathcal{R}_2$. These can be evaluated in the WKB approximation. The effective Hamiltonian is [57]

$$H(x, y, p_x, p_y) = ax(e^{p_x} - 1) + \frac{\sigma xy}{1 + \sigma\tau x}\left(e^{p_y - p_x} - 1\right) + y\left(e^{-p_y} - 1\right) + \frac{x^2}{2}\left(e^{-p_x} - 1\right). \tag{149}$$

The three deterministic fixed points become $M_1 = (0,0,0,0)$, $M_2 = (2a,0,0,0)$ and $M_3 = (x^*, y^*, 0, 0)$. In its turn, the deterministic limit cycle is an exact time-periodic solution of the Hamilton equations with $p_x = p_y = 0$. There are also two fluctuational fixed points:
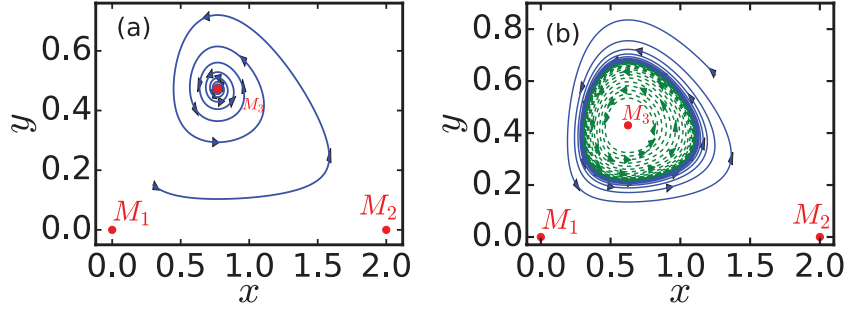
Figure 25. Deterministic trajectories of the Rosenzweig–MacArthur model (139) for $a = 1$, $\tau = 0.5$ and two different values of σ: $\sigma = 2.6$, where M_3 is a stable focus (a), and $\sigma = 3.2$, where M_3 is an unstable focus, and a stable limit cycle exists (b). Reprinted figure with permission from [57], copyright 2015 by the American Physical Society.

$$F_1 = (0, 0, 0, -\infty), \quad F_2 = \left[2a, 0, 0, \ln\left(\frac{1 + 2a\tau\sigma}{2a\sigma} \right) \right]. \tag{150}$$

The extinction rates $\mathcal{R}_1$ and $\mathcal{R}_2$ can be evaluated by evaluating the actions S_1 and S_2 along the instantons in the four-dimensional phase space. These instantons exit, at $t = -\infty$, the deterministic limit cycle and approach, at $t = \infty$, the fluctuational fixed point F_1 or F_2, respectively:

$$S_1 = \int_{(x_{lc}, y_{lc})}^{F_1} p_x \mathrm{d}x + p_y \mathrm{d}y \quad \text{and} \quad S_2 = \int_{(x_{lc}, y_{lc})}^{F_2} p_x \mathrm{d}x + p_y \mathrm{d}y,$$

The extinction rates $\mathcal{R}_1$ and $\mathcal{R}_2$ are, with exponential accuracy,

$$\mathcal{R}_1 \sim \exp(-NS_1), \quad \mathcal{R}_2 \sim \exp(-NS_2). \tag{151}$$

Until recently, no numerical algorithm existed for finding an instanton that describes a fluctuation-driven escape from a deterministically stable limit cycle. One such algorithm has been recently developed [57]. It is based on Floquet theory of differential equations with periodic coefficients, see e.g. [135]. The reader is referred to [57] for a detailed description of the algorithm. Examples of instantons, found with this algorithm, are shown in figure 26.

Figure 27 compares the extinction rates $\mathcal{R}_i$ obtained by solving numerically the (truncated) master equation (144), with the numerical result of the leading order WKB approximation, e^{-NS_i}. In its turn, figure 28 shows the MTE of foxes $\mathrm{MTE_F}$ and the ratio of the probabilities of the two extinction routes, $\mathcal{P}_1 / \mathcal{P}_2$, obtained by averaging over many Monte-Carlo simulations of the stochastic model. The same figure shows the corresponding leading-order WKB predictions $\left(\mathrm{e}^{-NS_1} + \mathrm{e}^{-NS_2} \right)^{-1} \simeq \mathrm{e}^{NS_2}$ and $\mathrm{e}^{N(S_2 - S_1)}$, respectively. In both cases a good agreement, up to undetermined WKB prefactors, was observed between the numerical and WKB results [57].

[57] also showed that the 'extinction action' changes in a non-analytic way as the system passes through the Hopf bifurcation of the deterministic theory at $\sigma = \sigma^*$, when the limit cycle is born. The second derivative of the action with respect to σ experiences a jump at $\sigma = \sigma^*$ as in a second-order phase transition. Finally, using the numerical results for the actions S_1 and S_2 along the instantons, and numerically solving the master equation (144), the authors of [57] determined the N-dependence of the prefactors $\mathcal{F}_i$ of the rates, $\mathcal{R}_i = \mathcal{F}_i \mathrm{e}^{-NS_i}$, $i = 1, 2$ at $N \gg 1$, and followed the change of the N-dependence as the system passes through the Hopf bifurcation. At $\sigma < \sigma^*$, the prefactors behave as $\mathcal{F} \propto N^{1/2}$. Above the bifurcation the prefactors are independent of N. These findings can be explained by simple normalization arguments [57].
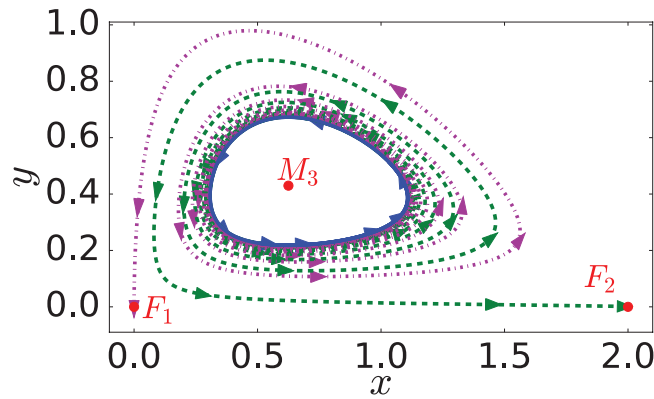
Figure 26. Extinction instantons (the x, y-projections) going from the stable deterministic limit cycle (solid line) to the fluctuational fixed points F_1 (dot-dashed) and F_2 (dashed). The parameters are $a = 1$, $\sigma = 3.2$, and $\tau = 0.5$. Reprinted figure with permission from [57], copyright 2015 by the American Physical Society.

8. Population extinction on heterogeneous networks

Population dynamics on heterogenous networks are very different from those in well-mixed settings because of the variability in the reaction rates along the network [136, 137]. While rare events have been studied on Erdös–Rényi networks, which are only slightly heterogenous, see e.g. [138], the recent years have witnessed a surge of works on the long-time population dynamics on degree-heterogenous networks with strong heterogeneity, such as scale-free networks [139]. [140–142] studied evolutionary models where the network nodes represent either a wild-type individual A, or a mutant B, and their interactions such as $A + B \to A + A$ or $A + B \to B + B$ determine the evolutionary dynamics. The main questions of interest here are in calculating the fixation probability of each species and the mean fixation time. The authors of [140–142] showed that the non-trivial network topology of heterogenous graphs, and especially of scale-free networks, may cause a dramatic decrease in the mean time of rare events to occur. A striking manifestation of this fact is that, in contrast to well-mixed systems (or complete graphs), the logarithm of the mean time to fixation may scale sublinearly with the system size N, $\ln \tau \sim N^\alpha$, where $\alpha < 1$.

Recently, in [77] the authors have investigated the long-time behavior and the MTE of an infectious disease on networks, using the classic SIS model of epidemiology, see equation (129). They employed the real-space WKB method to analyze the multivariate master equation describing the stochastic population dynamics on the network. We will now present their work in some detail.

Following [77], we assume an uncorrelated random network with a given degree distribution g_k, such that $\sum_k g_k = 1$, where the degree k is the number of links of a node. Simple graphs with N nodes can be generated from g_k in several ways, for example, with a configuration model network [137]. It is useful to represent such networks by the adjacency matrix A, where A_{ij} is 1 if nodes i and j are linked, and is 0 otherwise. An infected node is denoted by $\nu_i = 1$, and a susceptible node is denoted by $\nu_i = 0$. The SIS dynamics involve infection—a change of ν_i from 0 to 1—with probability per unit time $\beta(1 - \nu_i) \sum_j A_{ij} \nu_j$, and recovery—a change of ν_i from 1 to 0—with probability per unit time $\alpha \nu_i$, where β and α are the infection and recovery rates, respectively [77].
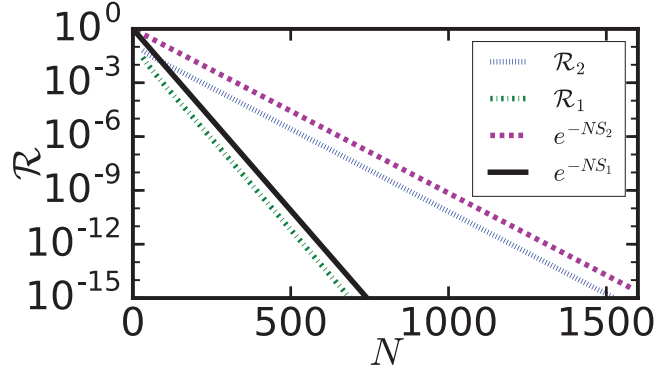
Figure 27. Extinction rates $\mathcal{R}_i$, determined by numerically solving the master equation (144) for different N, are compared to the extinction rates e^{-NS_i} computed in the leading-order WKB approximation. The parameters are $a = 1$, $\tau = 0.5$, and $\sigma = 3.1$. The actions along the instantons are $S_1 \simeq 0.0466$ and $S_2 \simeq 0.0211$. The observed vertical shifts are due to the undetermined WKB prefactors $\mathcal{F}_1$ and $\mathcal{F}_2$. Reprinted figure with permission from [57], copyright 2015 by the American Physical Society.
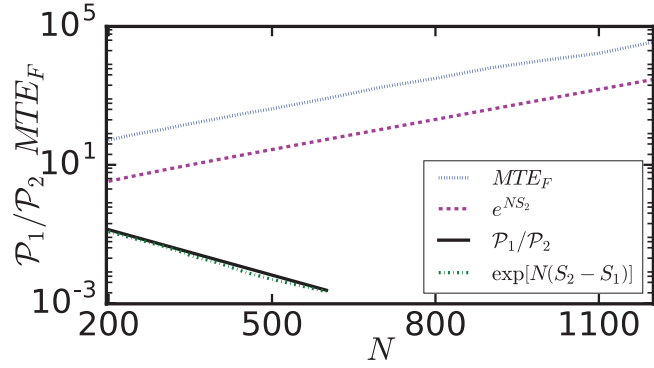

Figure 28. N-dependence of the MTE of foxes $\mathrm{MTE_F}$ and of the ratio of probabilities of the two extinction scenarios $\mathcal{P}_1/\mathcal{P}_2$ for fixed $a = 1$, $\tau = 0.5$, and $\sigma = 4$. The results obtained by averaging over many Monte Carlo simulations are compared to the leading-order WKB predictions. The values are plotted on a semi-log scale. The actions are $S_1 \simeq 0.0167$ and $S_2 \simeq 0.006\,65$. As $S_2 < S_1$, the MTE of the foxes is approximated in the WKB theory as $\mathrm{MTE_F} \simeq \mathrm{e}^{NS_2}$. The vertical shifts are due to the undetermined WKB prefactors. Reprinted figure with permission from [57], copyright 2015 by the American Physical Society.

To have a closed-form description, the authors of [77] assumed that there are no correlations between the nodes and replaced A_{ij} by its expectation value in an ensemble of networks $\langle A_{ij} \rangle = k_i k_j / (N \langle k \rangle)$, in the limit of large N. This (in general, uncontrolled) assumption is known as the 'annealed' network approximation: a mean-field theory of heterogeneous networks [140]. Given this mean-field form, the state of the network can be described by the number of infected nodes with degree k, I_k, which has the corresponding reactions and rates: $I_k \to I_k + 1$ with rate of infection

$$w_k^{\mathrm{inf}}(\mathbf{I}) = \beta k (N_k - I_k) \sum_{k'} k' I_{k'} / (N \langle k \rangle),$$

and $I_k \to I_k - 1$ with recovery rate

$$w_k^{\text{rec}}(\mathbf{I}) = \alpha I_k,$$

where $\mathbf{I} = (I_1, I_2, \ldots, I_{k_{\max}})$, and $N_k = N g_k$ [77].

In the stochastic description one writes down a multivariate master equation for the probability $P(\mathbf{I}, t)$ to find a vector of infected nodes $\mathbf{I}$ at time t. For networks with a general degree distribution the master equation reads

$$\frac{\partial}{\partial t}P(\mathbf{I}, t) = \sum_k w_k^{\text{inf}}(\mathbf{I} - \mathbf{1}_k)P(\mathbf{I} - \mathbf{1}_k, t) - w_k^{\text{inf}}(\mathbf{I})P(\mathbf{I}, t)$$
$$+ w_k^{\text{rec}}(\mathbf{I} + \mathbf{1}_k)P(\mathbf{I} + \mathbf{1}_k, t) - w_k^{\text{rec}}(\mathbf{I})P(\mathbf{I}, t), \tag{152}$$

where $\mathbf{1}_k = (0_{k_1}, 0_{k_2}, \ldots, 1_k, \ldots, 0_{k_{\max}})$. Essentially, this master equation describes the dynamics of $k_{\max}$ interacting populations.

For large but finite networks, the population enters a long-lived endemic state. We set $P(\mathbf{I}, t) = \pi(\mathbf{I})e^{-t/\tau}$, where τ is the MTE of the disease, and use the WKB ansatz for the QSD, $\pi(\mathbf{x}) = e^{-NS(\mathbf{x})}$, where $\mathbf{x} = \mathbf{I}/N$. In the leading order of $N \gg 1$ this leads to a stationary Hamilton–Jacobi equation $H(\mathbf{x}, \mathbf{p}) = 0$ with the Hamiltonian [77]

$$H(\mathbf{x}, \mathbf{p}) = \sum_k \left[\beta k (g_k - x_k)(e^{p_k} - 1) \sum_{k'} \frac{k' x_{k'}}{\langle k \rangle} + \alpha x_k (e^{-p_k} - 1) \right], \tag{153}$$

where $\mathbf{p} = \partial_{\mathbf{x}} S$ are the momenta. The problem boils down to finding the instanton: the zero-energy activation trajectory of the Hamilton's equations

$$\dot{y}_k = \tilde{\beta} k (1 - y_k) e^{p_k} \sum_{k'} \frac{k' g_{k'}}{\langle k \rangle} y_{k'} - y_k e^{-p_k},$$

$$\dot{p}_k = \tilde{\beta} k \sum_{k'} \frac{k' g_{k'}}{\langle k \rangle} \left[y_{k'}(e^{p_k} - 1) - (1 - y_{k'})(e^{p'_k} - 1) \right] - e^{-p_k} + 1. \tag{154}$$

Here $y_k = x_k/g_k$ is the fraction of each degree class infected, $\tilde{\beta} = \beta/\alpha$, and time is rescaled: $t \to \alpha t$. The activation trajectory must exit, at $t = -\infty$, the endemic fixed point $y_k^* = [1 + 1/(Y\tilde{\beta}k)]^{-1}$ and $p_k = 0$ and enter, at $t = \infty$, the extinction fixed point $y_k = 0$ and $p_k^* = -\ln[1 + \tilde{\beta}k(1 - P)]$. The functions Y and P satisfy the conditions $(Y, P) \to (0, 1)$ when $R_0 \to 1$, and $(Y, P) \to (1, 0)$ as $R_0 \to \infty$, where $R_0 \equiv \tilde{\beta}\langle k^2 \rangle/\langle k \rangle$ is the basic reproduction number. The MTE can be evaluated as $\tau \sim e^S$, where $S = \sum_k g_k \int p_k \dot{y}_k \mathrm{d}t$ is the action along the activation trajectory. In [77] this problem was solved numerically, whereas analytical results were obtained close to bifuraction, $R_0 - 1 \ll 1$, and in the strong infection limit, $R_0 \gg 1$.

Future work will have to account for correlations between nodes and incorporate more general statistics of adjacency matrix A_{ij}.

9. Extinction of spatial populations

It has been long known [143], that migration of individuals plays a crucial role in many environments of relevance to population biology and epidemiology [127]. In sections 7.1.3 we reviewed some recent work on extinction of populations whose individuals can migrate on a network of 'patches'. Here we will focus on populations whose individuals migrate in space. For simplicity, we will limit ourselves to one spatial dimension. Following [60, 78], let us

consider a single population residing on a one-dimensional lattice of $N \gg 1$ sites (or habitat patches) labeled by index $i = 1, 2, \ldots, N$. The population size n_i at each site varies in time as a result of two types of Markov processes. The first set of processes involves on-site stochastic dynamics of birth-death type, with birth and death rates $\lambda(n_i)$ and $\mu(n_i)$, respectively, where $\mu(0) = 0$. As there is no creation of new individuals 'from vacuum', one also has $\lambda(0) = 0$. The second process is random and independent migration of each individual between neighboring sites with migration rate coefficient D_0. What happens at the edges of the refuge, $i = 1$ and $i = N$, needs to be specified separately.

Assuming $n_i \gg 1$ and neglecting fluctuations, one obtains a set of coupled deterministic rate equations

$$\dot{n}_i = \lambda(n_i) - \mu(n_i) + D_0(n_{i-1} + n_{i+1} - 2n_i). \tag{155}$$

Let the characteristic population size on a single site, predicted by a non-trivial steady-state solution of equation (155), scale as $K \gg 1$. This implies [35, 72] that, in the leading order of K, we can represent the birth and death rates as

$$\lambda(n_i) = \mu_0 K \bar{\lambda}(q_i) \quad \text{and} \quad \mu(n_i) = \mu_0 K \bar{\mu}(q_i), \tag{156}$$

where $q_i = n_i/K$ is the rescaled population size at site i, $\bar{\lambda}(q_i) \sim \bar{\mu}(q_i) = \mathrm{O}(1)$, and μ_0 is a characteristic rate coefficient. Now equation (155) can be rewritten as

$$\dot{q}_i = \mu_0 f(q_i) + D_0(q_{i-1} + q_{i+1} - 2q_i), \tag{157}$$

where $f(q_i) = \bar{\lambda}(q_i) - \bar{\mu}(q_i)$. When $D_0 \gg \mu_0$, one can introduce a continuous spatial coordinate x instead of the discrete index i and replace the discrete Laplacian in equation (157) by the continuous one [78]. This leads to the reaction–diffusion equation

$$\partial_t q = \mu_0 f(q) + D \partial_x^2 q, \tag{158}$$

where $D = D_0 h^2$ is the diffusivity, and h is the lattice spacing. The system size becomes $L = Nh$. Equation (158) has been the subject of numerous studies, see e.g. [127, 144]. Meerson and Sasorov [78] considered periodic, $q(x + L) = q(x)$, and absorbing, $q(0) = q(L) = 0$, boundary conditions. The absorbing boundaries model extremely harsh conditions outside of the refuge [60, 143]. Results for still another type of boundaries—reflecting walls at $x = 0$ and $x = L$—can be easily obtained from the results for periodic boundaries.

Spatial deterministic profiles of established populations are described by stable steady-state solutions $q = q(x) > 0$ of equation (158). They satisfy the ODE

$$Dq''(x) + \mu_0 f(q) = 0 \tag{159}$$

subject to the chosen spatial boundary conditions. The first integral of this equation,

$$\frac{D}{2\mu_0} (q')^2 + V(q) = \text{const}, \tag{160}$$

with effective potential $V(q) = \int_0^q f(\xi) \, \mathrm{d}\xi$, makes the problem solvable in quadratures and readily yields the phase portrait of possible steady states on the plane (q, q'). Importantly, the reaction–diffusion equation (158) is a gradient flow, $\partial_t q = -\delta \mathcal{F}/\delta q$, where

$$\mathcal{F}[q(x,t)] = \int_0^L \mathrm{d}x \, \left[-\mu_0 V(q) + (1/2) D(\partial_x q)^2 \right]. \tag{161}$$

Therefore, it describes a deterministic flow towards a minimum of the Ginzburg–Landau free energy $\mathcal{F}[q]$. This helps identify linearly stable and unstable x-dependent solutions, as they correspond to local minima and maxima of $\mathcal{F}[q]$, respectively [144].

Going back to the lattice formulation and reintroducing the stochasticity of the on-site processes and of the migration, one should deal with a master equation for the multivariate distribution of population sizes on different patches [60, 78]. Being interested in extinction of established populations in a finite region of space, one arrives at a *quasi-stationary* master equation [78]. Applying the WKB ansatz to the quasi-stationary equation, and assuming a high migration rate, one finally arrives at the following Hamilton's equations in the real space [78]:

$$\partial_t q = \mu_0 \left[\bar{\lambda}(q)e^p - \bar{\mu}(q)e^{-p} \right] + D \left[\partial_x^2 q - 2\partial_x (q\partial_x p) \right], \tag{162}$$

$$\partial_t p = -\mu_0 \left[\bar{\lambda}'(q)(e^p - 1) + \bar{\mu}'(q)(e^{-p} - 1) \right] - D \left[\partial_x^2 p + (\partial_x p)^2 \right], \tag{163}$$

Equivalent equations, up to a canonical transformation, were obtained in [60]. Without migration, these equations coincide with those for well-mixed populations, see equation (12). Without births and deaths, they describe a simple particular case—non-interacting random walkers—of the so called macroscopic fluctuation theory: a WKB theory of large deviations in diffusive lattice gases [145].

The boundary conditions in space are problem-dependent, see above. For periodic or no-flux boundaries $p(x, t)$ is also periodic or no-flux at the boundaries. For the absorbing boundaries p vanishes at the boundary (up to corrections linear in the lattice constant) [78]. As shown in [78], the optimal path of a quasi-stationary population to extinction is again an instanton: a 'heteroclinic trajectory' between two 'fixed points' in the infinite-dimensional functional phase space of the system. The first 'fixed point' corresponds to the long-lived quasi-stationary distribution of the population size in space. In the absence of the Allee effect, the second 'fixed point' corresponds to a zero-population-size state with a nontrivial spatial profile of the conjugate momentum p. This extinction instanton can be found numerically by the Chernykh-Stepanov iteration algorithm [114], and in some limits analytically, see [60, 78].

In the presence of a sufficiently strong Allee effect the second 'fixed point' of the heteroclinic trajectory is *deterministic*. It describes the *critical nucleus* for extinction: [78]. Let us assume periodic boundaries, see figures 29 and 30. Here the only linearly stable nontrivial deterministic steady-state solution is the x-independent solution $q = q_2$, where $q_2 > 0$ is the attracting point of the deterministic equation $\dot{q} = \mu_0 f(q)$. A sufficiently large perturbation, however, triggers a *deterministic* transition from $q = q_2$ to the trivial solution $q = 0$ that is also linearly stable. The critical nucleus is an x-dependent deterministic solution $q_c(x)$ of equation (159) which is linearly unstable under the dynamics of equation (158). The critical nucleus corresponds to a local *maximum* of free energy (161) [144]. A small perturbation around the critical nucleus drives the system either to $q = 0$ or to $q = q_2$. In a finite-size system the critical nucleus depends on the system size L and corresponds to a phase trajectory inside the internal separatrix shown in figure 29(b) [78]. The critical nucleus exists only for sufficiently large systems, $L > L_c$, where the critical system size L_c can be obtained by linearizing equation (159) around $q = q_2$. At $L \gg L_c$ the critical nucleus approaches the internal separatrix in figure 29(b). If $f(q)$ is a cubic polynomial, the critical nucleus can be found analytically, otherwise it can be found numerically. Figure 30 shows the critical nuclei for two different values of $L > L_c$.

For extinction to occur a large fluctuation of the size of a stochastic population residing around $q = q_2$ only needs to create the critical nucleus. The further population dynamics toward extinction proceed essentially deterministically. What happens at $L \gg L_c$, see figure 30(b), is intuitively clear. Once having passed the critical nucleus, the solution $q(x, t)$ of equation (158) develops, on a fast time scale $\sim \mu_0^{-1}$, two outgoing deterministic 'extinction
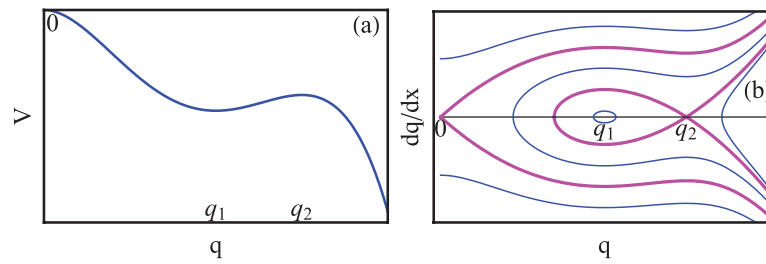
Figure 29. Effective potential $V(q)$ (a) and phase portrait (q, q') (b) for steady-state solutions of equation (158) for a strong Allee effect, $V(0) > V(q_2)$. Reprinted figure with permission from [78], copyright 2015 by the American Physical Society.

fronts' that drive the whole population to extinction on a time scale $\sim L/(\mu_0 D)^{1/2}$. As a result, the MTE is determined by the mean creation time of the critical nucleus. This quantity does not include an exponential dependence on the system size L (unless L is exponentially large in K [78]) and is therefore much shorter than the MTE in the absence of Allee effect.

For a *very* strong Allee effect—when the system is close to the saddle-node bifurcation, corresponding to the appearance of the stable fixed point $q = q_2$ of the deterministic equation $\dot{q} = \mu f(q)$—the WKB formulation of the population extinction problem turns out to be completely integrable, similarly to the complete integrability of the WKB problem for population explosion close to the saddle-node bifurcation, earlier studied in [60]. This integrability can be explained by establishing a direct mapping between this problem and the overdamped limit of the theory of homogeneous nucleation due to Langer [146]. We refer the reader to papers [60, 78] for details, derivations and additional examples.

Nucleation via fluctuations also plays a critical role in the formation of macroscopic *clusters* (for example, above the critical point in certain lattice gas models). This fundamental problem is of interest in several biological contexts such as quorum sensing and clustering of motile adhesive cancer cells [147]. In the latter case, clustering via nucleation can explain growth of recurrent tumors. Khain *et al* [147] studied this problem numerically and via a WKB approximation to a phenomenological master equation.

10. Large velocity fluctuations of population invasion fronts

Biological invasions is one of the key issues in contemporary ecology [148]. The term describes a broad variety of phenomena related to introduction and spread of non-native species in ecosystems. Biological invasions often have a dramatic negative impact on native eco-communities, and are considered as one of the main reasons for biodiversity loss in the world. They may also jeopardize agriculture and lead to economic losses. Mathematical modeling plays an important role in the understanding of biological invasions [9, 149]. Here we will consider simple one-population invasion fronts that, at the level of deterministic theory, are described by a traveling front solution of a reaction–diffusion equation. When the demographic noise is taken into account, the reaction front position fluctuates. What is the probability that the fluctuating reaction front moves slower or faster than its deterministic counterpart? The WKB approximation can be very useful in providing insightful answers. This theory assumes many particles in the transition region of the front. As it turns out, details crucially depend on whether the reaction front describes propagation into a state that is deterministically unstable or stable. We will only consider here fronts propagating into an unstable state, where the effects of noise are remarkably strong.
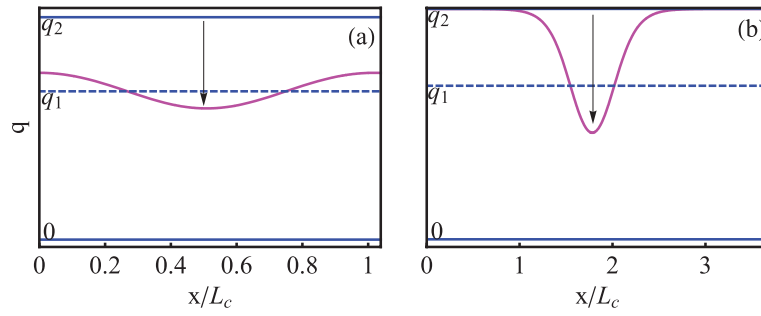
Figure 30. Linearly stable states $q = q_2$ and $q = 0$, linearly unstable state $q = q_1$ and critical nucleus for a strong Allee effect and periodic boundary conditions. The system size is weakly (a) and strongly (b) supercritical. The thick curves show the critical nucleus. The arrows indicate transitions, driven by rare large fluctuations and leading to a quick population extinction. Reprinted figure with permission from [78], copyright 2015 by the American Physical Society.

A celebrated example of a reaction front propagating into an unstable state is provided by the Fisher–Kolmogorov–Petrovsky–Piscounov (FKPP) equation [150],

$$\partial_t q = q - q^2 + \partial_x^2 q,$$
(164)

where $q(x, t)$ is the particle density. This equation describes invasion of an unstable state $q(x \to \infty, t) = 0$ by a stable state $q(x \to -\infty, t) = 1$. The FKPP equation is a fundamental model in population biology and mathematical genetics [150], but it also appears in many other applications. Equation (164) has a family of traveling-front solutions (TFSs) parameterized by their velocity c: $q(x,t) = Q_{0,c}(\xi)$, where $\xi = x - ct$. $Q_{0,c}(\xi)$ obeys the equation

$$Q_{0,c}'' + cQ_{0,c}' + Q_{0,c} - Q_{0,c}^2 = 0.$$
(165)

and boundary conditions $Q_{0,c}(-\infty) = 1$ and $Q_{0,c}(\infty) = 0$. The invasion front can have any velocity from the interval $c_0 \leqslant c < \infty$, where $c_0 = 2$. It is known, however, that for sufficiently steep initial conditions the solution of the time-dependent equation (164) approaches at long times a limiting TFS, $Q_{0,2}$, of equation (165) with the velocity $c_* = 2$ [150–152]. Importantly, the special velocity $c_0 = 2$ is determined only by the dynamics of the leading edge of the front, where equation (164) can be linearized around $q = 0$. That is, the full nonlinear front, described by equation (164), is 'pulled' by its leading edge, hence the term 'pulled fronts' [152], of which the FKPP equation (164) is the best known example.

The intrinsic (demographic) noise, unaccounted for by equation (164), causes deviations of the front position with time. These include a systematic front velocity shift and fluctuations that *typically* are diffusion-like. If $N \gg 1$ is the characteristic number of particles in the front region, the front velocity shift scales as $(\ln N)^{-2}$ [153–155], whereas the front diffusivity scales as $(\ln N)^{-3}$ [155–157]. The logarithmic N-dependencies are markedly different from the more expected $1/N$ dependencies of the same quantities for the fluctuating fronts propagating into *locally stable* states [79, 158–161]. This fact makes the pulled fronts much more interesting.

The starting point of the stochastic treatment of population invasion fronts can be the same as in the spatial extinction problem, see section 9. One starts from the exact master equation in the lattice formulation. Then, assuming a large N, one can apply a WKB ansatz directly to the time-dependent multivariate distribution of population sizes. Neglecting the second derivatives of the action and assuming a fast diffusion in space, one again arrives at the continuous Hamilton's equations (163) [78, 79]. Meerson and Sasorov [78] studied (negative) velocity

fluctuations of a pulled front on the example of reversible reactions $A \rightleftarrows 2A$ and independent random walk. In this case the (properly rescaled) equations (163) become

$$\partial_t q = q e^p - q^2 e^{-p} + \partial_x^2 q - 2\partial_x (q\partial_x p),$$ (166)

$$\partial_t p = 1 - e^p - 2q(e^{-p} - 1) - \partial_x^2 p - (\partial_x p)^2.$$ (167)

The specifics of the problem of fluctuations of the front position are reflected in the boundary conditions in space and in time [78, 81]. At $x \to -\infty$ there is a stationary distribution of the particle density, peaked at $q = 1$. Therefore, the boundary conditions at $x \to -\infty$ are $q(-\infty, t) = 1$ and $p(-\infty, t) = 0$ which correspond to the fixed point ($q = 1$, $p = 0$) of the *on-site* Hamiltonian $H_0(q, p) = q(e^p - 1) + q^2(e^{-p} - 1)$. The boundary condition at $x \to \infty$ is $q(\infty, t) = 0$. The momentum p can be unbounded at $x = \infty$, but it must be bounded at finite x [78]. The boundary conditions in time involve two kink-like profiles of q at $t = 0$ and $t = T \gg 1$ which are at distance $X \gg 1$ apart (where X can be positive, negative or zero). The front velocity is defined as $c = X/T$. A natural constraint in the context of invasion is that the initial particle density is zero to the right of some finite point in space.

Equations (166) and (167) look simpler in the new variables $Q = q e^{-p}$ and $P = e^p - 1$. But a crucial simplification appears when one realizes that, except in narrow boundary layers in time at $t = 0$ and $t = T \gg 1$, the optimal path of the system approaches a traveling-front solution (TFS), $Q(x, t) = Q(x - ct)$ and $P(x, t) = P(x - ct)$, of equations (166) and (167). As a result, these PDEs become ODEs. The probability density of observing the front velocity c is, in the leading order,

$$-\ln \mathcal{P}(X, T) = NT \, \mathcal{F}(X/T) = NT \, \mathcal{F}(c), \quad |c| < 2.$$ (168)

The optimal TFS, and the large deviation function (or rescaled action) $\mathcal{F}(c)$,

$$\mathcal{F}(c) = \int_{-\infty}^{\infty} d\xi \left[Q(\xi) - Q^2(\xi) \right] P^2(\xi),$$ (169)

can be easily found numerically [78]. Notice also, that the ODEs have an integral of motion that can be found explicitly [78]. Figure 31 shows the numerical solution for $c = 3/2$. As one can see, for this unusually slow front the P-front is well ahead of the Q-front. Essentially, the P-front modifies the effective reaction rates so as to slow down the front in comparison to its deterministic counterpart at $P = 0$. Figure 32(a) shows the numerically evaluated $\mathcal{F}(c)$, see equation (169), at different c.

One interesting regime where an analytic solution—by a matched asymptotic expansion—can be obtained is $c_0 - c \ll 1$ [78]. In this case the main contribution to the large deviation function comes from the leading edge of the front, $\xi \gg 1$, where $Q \ll 1$. As $|P| \sim 1$ there, this theory must be nonlinear. On the other hand, $|P| \ll 1$ in the left region, whereas $Q \ll 1$ in the right region. An analytical theory is possible because there is a *joint* region, where the strong inequalities $|P| \ll 1$ and $Q \ll 1$ hold simultaneously. We refer the reader for details to [78]. Finally one obtains

$$\mathcal{F}(c) \simeq 0.0074 \, \exp\left(-\frac{\pi}{\sqrt{2-c}} \right).$$ (170)

As a result,

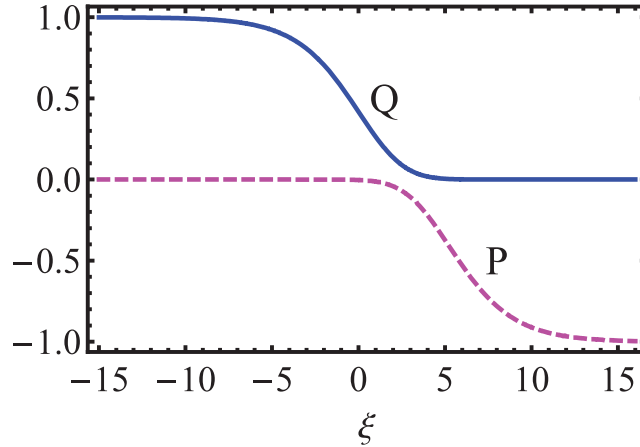$$\ln \mathcal{P}(X, T) \simeq -0.0074 \, NT \exp\left(-\frac{\pi}{\sqrt{2-c}} \right).$$ (171)

Figure 31. Numerical $Q(\xi)$ and $P(\xi)$ profiles for $c = 1.5$ obtained in [78]. Reprinted figure with permission from [80], copyright 2015 by the American Physical Society.

As one can see, the negative tail of the front speed distribution around c_0 is strongly non-Gaussian, and rapidly falls as c goes down. It also strongly depends on N. The latter property reflects the fact that, in order to considerably slow down the invasion front, a *multi-particle* optimal fluctuation is required. Being a WKB asymptotic, equation (171) holds only when there are many particles at the leading edge of the front, where $\xi \simeq \xi_0 \simeq \pi/\sqrt{2-c}$. This condition leads to the strong inequality

$$c_* - c \gg 2\pi^2 \ln^{-3} N, \tag{172}$$

where $c_* = 2 - \pi^2 \ln^{-2} N$ describes the *systematic* correction to the front velocity, coming from the noise [153]. By putting $c = c_* - \delta c$, where $\delta c \ll \ln^{-2} N$, one can see that the linear dependence on N in equation (171) cancels out, giving way to a much slower logarithmic dependence, compatible with a phenomenological theory of typical fluctuations of a pulled front's position [157].

It was realized in [78] that the result (171) holds, up to a c-independent numerical coefficient, for all pulled reaction fronts, where $A \to 2A$ is the only first-order birth process. It also holds, close to the transcritical bifurcation, for a whole class of reactions. The (properly rescaled) on-site Hamiltonian for this class of models, $H_0(Q, P) = QP(P - Q + 1)$ [106], was discussed in sections 6.2 and 7.1.1. Fluctuations of the front velocity for this class of models were studied in [81]. In particular, the authors solved numerically the complete time-dependent WKB equations—two coupled PDEs similar to equations (167). As they observed, the TFS is indeed the true optimal history of the system, except in the narrow boundary layers at $t = 0$ and $t = T \gg 1$, and it yields the leading-order contribution to $\ln \mathcal{P}$ as described by the scaling relation (168).

Large *positive* deviations of the pulled front velocity have an entirely different nature. In this case the distribution $\mathcal{P}(X, T)$ is independent of N in the leading order of theory [162, 163]. The positive deviations are of a non-WKB character, as they are determined by only a few particles that run ahead of the front.
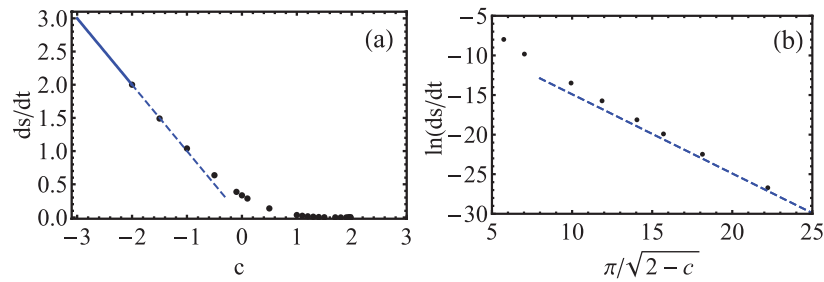
Figure 32. (a) Symbols: the large deviation function $\mathcal{F}(c)$, see equation (169), found numerically in [78], versus the front velocity c. Straight line: the asymptote $\mathcal{F}(c) = -c$ which becomes exact at $c \leqslant -2$ [78]. (b) $\ln \mathcal{F}(c)$ versus $\pi/\sqrt{2-c}$ for $c \geqslant 1.5$. Symbols: numerical results. Dashed line: asymptote (170). Reprinted figure with permission from [80], copyright 2015 by the American Physical Society.

11. Summary and some open questions

We have reviewed recent progress in applying the WKB approximation to the study of large deviations in stochastic population dynamics. The WKB approximation yields a versatile and universal framework for analysis of large deviations. In many cases it enables one to determine the optimal path to the specified large deviation and to evaluate, analytically or numerically, the mean time to extinction/fixation/switching, and/or the corresponding probabilities. Different variants of the WKB approximation, developed by physicists, are steadily 'invading' communities of mathematical biologists, and this process is likely to continue.

There are still many unresolved questions. Here are some of them. It would be interesting to search for additional instances of time-scale separation in multi-population systems in order to have a better analytical insight. Large deviations on heterogeneous networks remain largely unexplored beyond the mean-field-like 'annealed network' assumption. In spatially-extended large-deviation problems, time-scale separation seems to be the only hope for analytical progress, except when the optimal path is describable in terms of traveling wave solutions of the WKB equations, and the problem simplifies dramatically. Finally, it would be very interesting to investigate the effect of static disorder ('good' and 'bad' regions of space) on the large deviations in spatially-explicit populations.

Acknowledgments

References

[1] Delbrück M 1940 *J. Chem. Phys.* **8** 120
[2] Bartlett M S 1960 *Stochastic Population Models in Ecology and Epidemiology* (New York: Wiley)

[3]   Karlin S and McGregor J 1964 *Stochastic Models in Medicine and Biology* (Madison, WI: University of Wisconsin Press)
[4]   Bailey N T J 1975 *The Mathematical Theory of Infectious Diseases and its Applications* (London: Griffin)
[5]   Nisbet R M and Gurney W S C 1982 *Modelling Fluctuating Populations* (New York: Wiley)
[6]   Nåsell I 1985 *Hybrid Models of Tropical Infections* (Berlin: Springer)
[7]   Gabriel J P, Lefever C and Picard P (ed) 1990 *Stochastic Processes in Epidemic Theory* (Berlin: Springer)
[8]   Andersson H and Britton T 2000 *Stochastic Epidemic Models and Their Statistical Analysis* (New York: Springer)
[9]   Allen L 2003 *An Introduction to Stochastic Processes with Applications to Biology* (Upper Saddle River, NJ: Pearson Education)
[10] Nåsell I 2011 *Extinction and Quasi-stationarity in the Stochastic Logistic SIS Model* (Berlin: Springer)
[11] Tsimring L S 2014 *Rep. Prog. Phys.* **77** 026601
[12] Allen J 2015 *Stochastic Population and Epidemic Models—Persistence and Extinction* (Berlin: Springer)
[13] Spagnolo B, Valenti D and Fiasconaro A 2004 *Math. Biosci. Eng.* **1** 185
[14] Ovaskainen O and and Meerson B 2010 *Trends Ecol. Evol.* **25** 643
[15] Weber M F and Frey E 2017 *Rep. Prog. Phys.* **80** 046601
[16] Bressloff P C 2017 *J. Phys. A: Math. Theor.* **50** 133001
[17] Ott E 2002 *Chaos in Dynamical Systems* (Cambridge: Cambridge University Press)
[18] Grassberger P 1989 *J. Phys. A: Math. Gen.* **22** 3283
[19] Graham R, Hamm A and Tel T 1991 *Phys. Rev. Lett.* **66** 3089
[20] Khovanov I A, Luchinsky D G, McClintock P V and Silchenko A N 2008 *Int. J. Bifurcation Chaos* **18** 1727
[21] Landau L D and Lifshitz E M 1991 *Quantum Mechanics* (Oxford: Pergamon)
[22] Freidlin M I and Wentzell A D 1984 *Random Perturbations of Dynamical Systems* (New York: Springer)
[23] Dykman M I and Krivoglaz M A 1984 *Soviet Physics Reviews* vol 5, ed I M Khalatnikov (New York: Harwood Academic) pp 265–441
[24] Graham R 1989 *Theory of Continuous Fokker–Planck Systems* ed F Moss and P McClintock (Cambridge: Cambridge University Press) ch 7 p 225
[25] Kubo R, Matsuo K and Kitahara K 1973 *J. Stat. Phys.* **9** 51
[26] Knessl C, Matkowsky B J, Schuss Z and Tier C 1985 *SIAM J. Appl. Math.* **46** 1006
[27] Dykman M I, Mori E, Ross J and Hunt P M 1994 *J. Chem. Phys.* **100** 5735
[28] Roma D M, O'Flanagan R A, Ruckenstein A E, Sengupta A M and Mukhopadhyay R 2005 *Phys. Rev.* E **71** 011902
[29] Doering C R, Sargsyan K V, Sander L M and Vanden-Eijnden E 2007 *J. Phys.: Condens. Matter* **19** 065145
[30] Kessler D A and Shnerb N M 2007 *J. Stat. Phys.* **127** 861
[31] Dykman M I, Schwartz I B and Landsman A S 2008 *Phys. Rev. Lett.* **101** 078101
        Schwartz I B, Billings L, Dykman M and Landsman A 2009 *J. Stat. Mech.* P01005
[32] Meerson B and Sasorov P V 2008 *Phys. Rev.* E **78** 060103
[33] Escudero C and Kamenev A 2009 *Phys. Rev.* E **79** 041149
[34] Khasin M and Dykman M I 2009 *Phys. Rev. Lett.* **103** 068101
[35] Assaf M and Meerson B 2010 *Phys. Rev.* E **81** 021116
[36] Mobilia M and Assaf M 2010 *Europhys. Lett.* **91** 10002
[37] Assaf M and Mobilia M 2010 *J. Stat. Mech.* P09009
[38] Khasin M, Dykman M I and Meerson B 2010 *Phys. Rev.* E **81** 051925
[39] Bressloff P C 2010 *Phys. Rev.* E **82** 051903
[40] Assaf M and Mobilia M 2011 *J. Theor. Biol.* **275** 93
[41] Lohmar I and Meerson B 2011 *Phys. Rev.* E **84** 051901
[42] Black A J and McKane A J 2011 *J. Stat. Mech.* P12006
[43] Assaf M, Roberts E and Luthey-Schulten Z 2011 *Phys. Rev. Lett.* **106** 248102
[44] Gottesman O and Meerson B 2012 *Phys. Rev.* E **85** 021140
[45] Newby J M 2012 *Phys. Biol.* **9** 026002
[46] Black A J, Traulsen A and Galla T 2012 *Phys. Rev. Lett.* **109** 028101

J. Phys. A: Math. Theor. **50** (2017) 263001

Topical Review

[47] Khasin M, Meerson B, Khain E and Sander L M 2012 *Phys. Rev. Lett.* **109** 138104
[48] Gabel A, Meerson B and Redner S 2013 *Phys. Rev.* E **87** 010101
[49] Earnest T M, Roberts E, Assaf M, Dahmen K and Luthey-Schulten Z 2013 *Phys. Biol.* **10** 026002
[50] Assaf M, Roberts E, Luthey-Schulten Z and Goldenfeld N 2013 *Phys. Rev. Lett.* **111** 058102
[51] Vardi N, Levy S, Assaf M, Carmi M and Barkai N 2013 *Curr. Biol.* **23** 2051
[52] Assaf M, Mobilia M and Roberts E 2013 *Phys. Rev. Lett.* **111** 238101
[53] Roberts E, Be'er S, Bohrer C, Sharma R and Assaf M 2015 *Phys. Rev.* E **92** 062717
[54] Be'er S, Assaf M and Meerson B 2015 *Phys. Rev.* E **91** 062126
[55] Newby J M 2015 *J. Phys. A: Math. Theor.* **48** 185001
[56] Méndez V, Assaf M, Campos D and Horsthemke W 2015 *Phys. Rev.* E **91** 062133
[57] Smith N R and Meerson B 2016 *Phys. Rev.* E **93** 032109
[58] Gardiner C W 2004 *Handbook of Stochastic Methods* (Berlin: Springer)
[59] van Kampen N G 2001 *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland)
[60] Elgart V and Kamenev A 2004 *Phys. Rev.* E **70** 041106
[61] Assaf M and Meerson B 2006 *Phys. Rev.* E **74** 041115
[62] Assaf M and Meerson B 2006 *Phys. Rev. Lett.* **97** 200602
[63] Assaf M and Meerson B 2007 *Phys. Rev.* E **75** 031122
[64] Escudero C and Rodríguez J 2008 *Phys. Rev.* E **77** 011130.
[65] Kamenev A and Meerson B 2008 *Phys. Rev.* E **77** 061107
[66] Assaf M, Kamenev A and Meerson B 2008 *Phys. Rev.* E **78** 041123
[67] Assaf M, Kamenev A and Meerson B 2009 *Phys. Rev.* E **79** 011127
[68] Assaf M, Meerson B and Sasorov P V 2010 *J. Stat. Mech.* P07018
[69] Be'er S, Heller-Algazi M and Assaf M 2016 *Phys. Rev.* E **93** 052117
[70] Be'er S and Assaf M 2016 *J. Stat. Mech.* P113501
[71] Gaveau B, Moreau M and Toth J 1996 *Lett. Math. Phys.* **37** 285
[72] Doering C R, Sargsyan K V and Sander L M 2005 *Multiscale Modelling Simul.* **3** 283
[73] Meerson B and Sasorov P V 2009 *Phys. Rev.* E **80** 041130
[74] Kamenev A, Meerson B and Shklovskii B 2008 *Phys. Rev. Lett.* **101** 268103
[75] Levine E Y and Meerson B 2013 *Phys. Rev.* E **87** 032127
[76] Bacaër N 2015 *J. Math. Biol.* **71** 491
[77] Hindes J and Schwartz I B 2016 *Phys. Rev. Lett.* **117** 028302
[78] Meerson B and Sasorov P V 2011 *Phys. Rev.* E **83** 011129
[79] Meerson B, Sasorov P V and Kaplan Y 2011 *Phys. Rev.* E **84** 011147
[80] Meerson B and Sasorov P V 2011 *Phys. Rev.* E **84** 030101
[81] Meerson B, Vilenkin A and Sasorov P V 2013 *Phys. Rev.* E **87** 012117
[82] Nåsell I 2001 *J. Theor. Biol.* **211** 11
[83] Nowak M A 2006 *Evolutionary Dynamics* (Harvard, MA: Harvard University Press)
[84] Moran P A P 1962 *The Statistical Processes of Evolutionary Theory* (Oxford: Clarendon Press)
[85] Altland A, Fischer A, Krug J and Szendro I G 2011 *Phys. Rev lett.* **106** 088101
[86] Allee W C, Emerson A E, Park O and Park T 1949 *Principles of Animal Ecology* (Philadelphia, PA: Saunders)
[87] Böttger K, Hatzikirou H, Voss-Böhme1 A, Cavalcanti-Adam E A, Herrero M A and Deutsch A 2016 *PLOS Comput. Biol.* **11** e1004366
[88] Ptashne M 1992 *A Genetic Switch: Phage Lambda and Higher Organisms* (Cambridge, MA: Blackwell Scientific)
[89] Roberts E, Magis A, Ortiz J O, Baumeister W and Luthey-Schulten Z 2011 *PLoS Comput. Biol.* **7** e1002010
[90] Bressloff P C 2014 *Stochastic Processes in Cell Biology* vol 41 (Heidelberg: Springer)
[91] Mehta P, Mukhopadhyay R and Wingreen N S 2008 *Phys. Biol.* **5** 026005
[92] Gardner T S, Cantor C R and Collins J J 2000 *Nature* **403** 339–42
[93] Allen R J, Warren P B and Ten Wolde P R 2005 *Physi. Rev. Lett* **94** 018104
[94] Lipshtat A, Loinger A, Balaban N Q and Biham O 2006 *Phys. Rev. Lett.* **96** 188101
[95] Biancalani T and Assaf M 2015 *Phys. Rev. Lett.* **115** 208101
[96] Shahrezaei V and Swain P S 2008 *Proc. Natl Acad. Sci.* **105** 17256–61
[97] Paulsson J and Ehrenberg M 2000 *Phys. Rev. Lett.* **84** 5447
[98] Nelson D L and Cox M M 2005 *Lehninger Principles of Biochemistry* 4th edn (New York: Freeman)
[99] Pearson J E, Krapivsky P and Perelson A S 2011 *PLoS Comput. Biol.* **7** e1001058

[100] González-Suárez M and Revilla E 2013 *Ecol. Lett.* **16** 242
[101] Be'er S and Assaf M 2016 in preparation
[102] Turner J W and Malek-Mansour M 1978 *Physica* A **93** 517
[103] Dykman M I, Rabitz H, Smelyanskiy V N and Vugmeister B E 1997 *Phys. Rev. Lett.* **79** 1178
[104] Dykman M I, Golding B, McCann L I, Smelyanskiy V N, Luchinsky D G, Mannella R and McClintock P V E 2001 *Chaos* **11** 587
[105] Landau L D and Lifshitz E M 1976 *Mechanics* (Oxford: Pergamon)
[106] Elgart V and Kamenev A 2006 *Phys. Rev.* E **74** 041101
[107] Leigh E G 1981 *J. Theor. Biol.* **90** 213
[108] Lande R 1993 *Am. Naturalist* **142** 911
[109] Elowitz M B, Levine A J, Siggia E D and Swain P S 2002 *Science* **297** 1183
[110] Assaf M and Meerson B 2008 *Phys. Rev. Lett.* **100** 058105
[111] Hethcote H W 1976 *Math. Biosciences* **28** 335
      Hethcote H W 2000 *SIAM Rev.* **42** 599
      Jacquez J A and Simon C P 1993 *Math. Biosciences* **117** 77
      Nåsell I 1999 *J. R. Stat. Soc.* B **61** 309
[112] van Herwaarden O A and Grasman J 1995 *J. Math. Biol.* **33** 581
[113] van Herwaarden O A 1997 *J. Math. Biol.* **35** 793
[114] Chernykh A I and Stepanov M G 2001 *Phys. Rev.* E **64** 026306
[115] Gurarie V and Migdal A 1996 *Phys. Rev.* E **54** 4908
[116] Balkovsky E, Falkovich G, Kolokolov I and Lebedev V 1997 *Phys. Rev. Lett.* **78** 1452
[117] Grafke T, Grauer R and Schäfer T 2015 *J. Phys. A: Math. Theor.* **48** 333001
[118] Wang Z, Bauch C T, Bhattacharyya S, d'Onofrio A, Manfredi P, Matjaž P, Perra N, Salathé M and Zhao D 2016 Phys. Rep. **664** 1–114
[119] Lotka A J 1920 *Proc. Natl Acad. Sci.* 6 410
      Volterra V 1926 *Nature* **118** 558
[120] Khasin M, Meerson B and Sasorov P V 2010 *Phys. Rev.* E **81** 031126
[121] Eriksson A, Elias-Wolff F and Mehlig B 2013 *Theor. Popul. Biol.* **83** 101
[122] Choi P J, Cai L, Frieda K and Xie X S 2008 *Science* **322** 442
[123] Elton C and Nicholson M 1942 *J. Animal Ecol.* **11** 215
[124] Butler L 1953 *Canadian J. Zool.* **31** 242
[125] Bazykin A D 1998 *Nonlinear Dynamics of Interacting Populations* (Singapore: World Scientific)
[126] Odum E and Barrett G W 2004 *Fundamentals of Ecology* (Boston, MA: Cengage Learning)
[127] Murray J D 2008 *Mathematical Biology: I. An Introduction* (New York: Springer)
[128] Rosenzweig M L and MacArthur R H 1963 *Am. Naturalist* **97** 209
[129] Anderson R and May R M 1991 *Infectious Diseases of Humans: Dynamics and Control* (Oxford: Oxford University Press)
[130] Kirschner D and Panetta J C 1998 *J. Math. Biol.* **37** 235
[131] D'Onofrio A and Gandolfi A 2009 *Math. Med. Biol.* **26** 63
[132] Schlögl F 1972 *Z. Phys.* **253** 147
[133] Cheng K S 1981 *SIAM J. Math. Anal.* **12** 541
[134] Smith H L 2008 The Rosenzweig–MacArthur predator-prey model (https://math.la.asu.edu/~halsmith/Rosenzweig.pdf)
[135] Ward M J 2010 Basic Floquet theory (http://www.emba.uvm.edu/~jxyang/teaching/Floquet_theory_Ward.pdf)
[136] Pastor-Satorras R and Vespignani A 2001 *Phys. Rev. lett.* **86** 3200
[137] Newman M E J 2010 *Networks: an Introduction* (New York: Oxford University Press)
[138] Lindley B S, Shaw L B and Schwartz I B 2014 *Europhys. Lett.* **108** 58008
[139] Barabási A-L and Albert R 1999 *Science* **286** 509
[140] Antal T, Redner S and Sood V 2006 *Phys. Rev. Lett.* **96** 188104
[141] Sood V, Antal T and Redner S 2008 *Phys. Rev.* E **77** 041121
[142] Assaf M and Mobilia M 2012 *Phys. Rev. Lett.* **109** 188701
[143] Skellam J G 1951 *Biometrika* **38** 196
[144] Mikhailov A S 1990 *Foundations of Synergetics I. Distributed Active Systems* (Berlin: Springer)
[145] Bertini L, De Sole A, Gabrielli D, Jona-Lasinio G and Landim C 2015 *Rev. Mod. Phys.* **87** 593
[146] Langer J S 1969 *Ann. Phys.* **54** 258
[147] Khain E, Khasin M and Sander L M 2014 *Phys. Rev.* E **90** 062702

[148] Williamson M 1996 *Biological Invasions* (Berlin: Springer)
[149] Petrovskii S V and Bai-Lian L 2006 *Exactly Solvable Models of Biological Invasion* (Boca Raton, FL: CRC Press)
[150] Fisher R A 1937 *Ann. Eugenics* **7** 355
    Kolmogorov A, Petrovsky I and Piscounov N 1937 *Bull. Univ. État Moscou* A **1** 1
[151] Bramson M 1983 *Mem. Am. Math. Soc.* **44** 285
[152] van Saarloos W 2003 *Phys. Rep.* **386** 29
[153] Brunet É and Derrida B 1997 *Phys. Rev.* E **56** 2597
[154] Pechenik L and Levine H 1999 *Phys. Rev.* E **59** 3893
[155] Brunet É and Derrida B 1999 *Comput. Phys. Commun.* **121–2** 376
    Brunet É and Derrida B 2001 *J. Stat. Phys.* **103** 269
[156] Panja D 2003 *Phys. Rev.* E **68** 065202
[157] Brunet É, Derrida B, Mueller A H and Munier S 2006 *Phys. Rev.* E **73** 056126
[158] Kessler D A, Ner Z and Sander L M 1998 *Phys. Rev.* E **58** 107
[159] Panja D 2004 *Phys. Rep.* **393** 87
[160] Khain E and Meerson B 2013 *J. Phys. A: Math. Theor.* **46** 125002
    Khain E and Meerson B 2013 *J. Phys. A: Math. Theor.* **46** 139601
[161] Rulands S, Klünder B and Frey E 2013 *Phys. Rev. Lett.* **110** 038102
[162] Derrida B and Shi Z 2016 *J. Stat. Phys.* **163** 1285
[163] Derrida B, Meerson B and Sasorov P V 2016 *Phys. Rev.* E **93** 042139