2. Probabilistic and Stochastic Modelling (Population Genetics)

Modelling biological systems can be tricky. Feedback loops operating across hugely disparate spatial and temporal scales give rise to the huge richness and complexity of the living world. The process of abstracting from reality to a model involves removing much of this intricate detail. The more detail that has been discarded, the more the model is said to be 'simple'. The more simple a model is, the better chance we have of analysing it with mathematics to make predictions, gain deep understanding and derive general principles. Mathematical biologists face the challenge of constructing models that are simple enough to be amenable to analysis, but that still retain the essential features of the biological system under study.

There is one particularly powerful abstraction that is performed routinely and often subconsciously: we replace quantities with our expectation of those quantities, often without regard for the error. This has entered the public consciousness recently with attention on the basic reproductive number of Covid-19. Any single transmission of the coronavirus is governed by a sequence of events across vastly different scales (sub-cellular to intercontinental) which is so complicated as to be completely unpredictable. Yet these personal stories are aggregated at a population level to produce a single number quantifying the rate of new infections and thus the exponential growth of the disease. By relaxing the assumption of perfect predictability and embracing stochastic thinking, mathematical biology can better model biological systems, can expand the breadth and sophistication of predictions made, and can expose the counter-intuitive ways that randomness brings emergent order to the natural world.

All biological systems are inherently stochastic[15]. Plant and animal cells are complicated and crowded environments, with myriad competing and complementary processes and events taking place. Births and deaths of organisms, which ultimately drive the dynamics of whole populations and the evolution of species, are unpredictable in their timing, causes, and consequences. Empirical scientists deal with these uncertainties using statistics. For mathematical modellers, the correct proxy is the use of random variables to replace unknowns. In this way uncertainty can be carried forward through the model in a quantifiable way []. There are broadly three different formalisms that can be used to model a stochastic system.

The first is to consider the dynamics probabilistically in a discrete state space. Consider a vector of $N+1$ discrete states. These states could, for instance, represent the number of bacterial cells on a plate. The probability of being in a state $n = 0, 1, 2, \ldots, N$ at time t, $P_n(t)$, is governed by the following equation,

[15] From the Greek *stokhastikos* meaning "guessable", we use the word stochastic here to account for uncertainty in the timing of and outcome of events.

known as Kolmogorov's forward equation or the master equation;

$$\frac{\mathrm{d}P_n(t)}{\mathrm{d}t} = \sum_{n' \neq n} \left[T(n|n')P_{n'}(t) - T(n'|n)P_n(t) \right] . \tag{2.0.1}$$

Here the functions $T(n|n')$ are known as the probability transition rates, which can be interpreted as the probability per unit time of a transition from state n' to a state n. The interpretation of the master equation is intuitively clear: the probability that the system is in state n increases with the probability that the system moves into it from one of the surrounding states n', but decreases with the probability that the system is already in state n but transitions to another state. The transition rates fully determine the model under consideration.

The second is to consider the dynamics probabilistically in a continuous state space. Consider a variable $x \in [0, l]$, which could, for instance, represent the abundance of a protein within a cell. The probability of finding the system in an interval $x \in [a, b]$ at time t, $\int_a^b p(x,t)dx$ is governed by the following advection-diffusion equation, known as Fokker-Planck equation (FPE);

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[A(x)p(x,t)\right] + \frac{\partial^2}{\partial x^2}\left[B(x)p(x,t)\right] . \tag{2.0.2}$$

The function $A(x)$ in the first "spatial" derivative governs 'bulk advection', while the $B(x)$ in the second derivative governs the diffusion of the probability density function $p(x, t)$. Note that although this PDE in two variables, x and t, is far more amenable to analysis than the $N + 1$ partial difference equations comprising the master equation in Eq. (2.0.1), it also comes with an implicit assumption about the underlying process being modelled, namely that the state space is continuous (probability flows to neighbouring states only). While this assumption may be appropriate for modelling some biological processes (e.g. population growth) it may be inappropriate for others (e.g. genetic switches).

Finally we could model the behaviour as a stochastic differential equation (SDE). The typical form in 1D is

$$\frac{\mathrm{d}x}{\mathrm{d}\tau} = A(x) + \frac{1}{\sqrt{N}}\eta(x,\tau) \tag{2.0.3}$$

where the function $A(x)$ can be crudely understood as controlling the "deterministic component" of the dynamics, and $\eta(x, \tau)$ is a series of "multiplicative" (x dependent) Gaussian white noise terms with zero mean and with a correlator $\langle\eta(x,\tau)\eta(x,\tau')\rangle = B(x)\delta(\tau - \tau')$. These can be viewed as giving the deterministic dynamics (governed by $\dot{x} = A(x)$) a series of random "kicks" along its trajectory, leading to a stochastic profile. Note that while Eqs. (2.0.1) and (2.0.2) describe the evolution of probability distributions, Eq. (2.0.3) describes the evolution of an individual trajectory. Each realisation of Eq. (2.0.3) will thus give a different result, but a probability distribution can be constructed by averaging over many of these trajectories.

While the descriptions in Eqs. (2.0.1-2.0.3) may seem very different, it turns out that there is a way of starting out with a model cast in terms of the master equation (sometimes called a microscopic description, as it takes account of discrete indviduals) to the FPE and SDE (sometimes called mesoscopic descriptions, as they include some of the unpredictability of the microscopic system). Indeed, the notation for $A(x)$ and $B(x)$ in Eqs. (2.0.2-2.0.3) reflect this. However one could simply propose an ad-hoc FPE (as was done in the early days of mathematical population genetics) or SDE (as is frequenctly done in finance). We will explore these mappings in Section 2.2. However before we do, some biological terminology must be introduced.

2.1. An introduction to population genetics. [16]

The genetic identity of any individual, be it bacteria or sequoia, is determined by its genome. This is a "list" of all the genetic information of an organism, encoded in a nucleotide sequence, DNA (or RNA for viruses). The deep biological details are beyond the scope of this course, but for our purposes it suffices to know that an individual's genome can be represented by a string of letters, A, C, G, and T, for DNA (or A, C, U, and G for RNA). These strings are packaged in discrete sections called chromosomes.

The number of chromosomes vary widely across organisms. Bacteria have a single chromosome[17]. Meanwhile the green single-celled algae *Chlamydomonas reinhardtii* has 17 chromosomes, and the yeast *S. cerevisiae* has 16. Each of these chromosomes may be of a different length, and they do not share similarities in their composition. Organisms featuring this single "set" of chromosomes are termed haploid.

Chromosomes can also be arranged in pairs. This organisation is termed diploidy. Humans are diploid, and so this may be a more familiar genetic system to those unfamiliar with genetics. Humans have 22 pairs of chromosomes, with the chromosomes in each pair having the same length and similar structure (they are homologous). The picture is complicated slightly in humans as they also feature a 23rd non-homologous (different sized) pair of chromosomes; the sex chromosomes, labelled X and Y. Females carry two X chromosomes, and males carry an X and a significantly shorter Y chromosome (1/3 the size of the X chromosome). The existance of these homologous pairs is an important factor in sexual reproduction and the consequent genetic recombination (shuffling of parental genes in progeny), which we will touch upon in Section **??**. Note that while most of the mammalian life cycle is diploid, there is one stage at which

[16] For mathematicians! This is very much an informal discussion aimed at getting you tooled up with the appropriate terminology to tackle the maths.

[17] The chromosome is actually circular, with the "list"of genetic information looping back on itself. However for our purposes it's more useful conceptually to continue to think of this as a large string

haploid cells are produced; gametes (egg and sperm). These sex cells are haploid, but with morphology determined by the diploid "parent" (i.e. sperm can carry either X or Y chromosomes).

Things can get more complicated still in plants, many of which have more than two sets of chromosomes. Common wheat (*Triticum aestivum*) has six sets of chromosomes (it's hexaploid) while commercial strawberries[18] have eight (they're octoploid). Strange things also happen in the eusocial insects (e.g. colony wasp, bee, and ant species), which feature both haploid and diploid individuals (fertilised eggs are diploid, and develop into females, while unfertilised eggs develop into males). This unusual genetic system is thought to contribute to their capacity for cooperation.

There is a rapidly increasing amount of information about the genomes of organisms on earth, with full sequences now available for many species. However from an applied mathematics perspective, we wish to simplify our modelling of how these sequences come about and any patterns that may emerge. With this in mind it is common to restrict ourselves to considering a single location[19] on the genome (i.e. a single location on a chromosome, or chromosome pair), and ignore the surrounding genetic information. We call such a location a locus. At this locus can be a number of variants on a particular chromosome (e.g. A, C, G, and T). We will call these variants alleles. However we will not typically label these alleles as A, C, G, and T, but rather A and B, or A and a. This is to preserve generality[20].

As some of these definitions might become clearer with a mathematical foundation, we begin with a simple example that we will exploit in later sections.

2.1.1. *The two-allele haploid Moran model with mutation: master equation*

Consider a single locus in a haploid population with two alleles, A and B. The number of individuals carrying each allele is given by the discrete numbers n_A and n_B respectively. A large class of population genetic models called Moran models assume a fixed population size $N = n_A + n_B$ and continuous time. In order to fix this population size, birth and death events must be coupled, so that each birth simultaneously leads to the death of another individual. In this way, the state of the system is determined by a single variable $n_a = n$ (where we have dropped the subscript for simplicity), since $n_b = N - n_a$. We will also allow mutations at

[18] The garden strawberry was first bred in Brittany, France, in the 1750s via a cross of *Fragaria virginiana* from eastern North America and *Fragaria chiloensis*, which was brought from Chile by Amedee-Francois Frezier in 1714. Prior to this most europeans ate only the wild european strawberry, which is delicious but tiny.

[19] Or small number of locations

[20] Although we refer to a locus as a single location, it could equally refer to a short sequence, such that we could have for intstance A represent the sequence CATGA and B represent the sequence CACGA at the same locus consisting of five nucleotides

a rate μ to transform individuals carrying allele A to individuals carrying allele B and vice-versa. For simplicity, we will here assume that these mutation events are independent from birth/death events.

The set-up of the problem (a discrete state space in continuous time) makes the problem ideally tackled with the master equation, Eq. (2.0.1). In order to define the master equation for this particular model, we need to define the probability transition rates, $T(n'|n)$. In this model, only birth (a type A reproduces, a type B dies) and death (a type B reproduces, a type A dies) events are possible, so for any given n, $n' = n+1$ (birth) or $n' = n-1$ (death). If individuals are picked randomly from the population, the transition rates are then

$$\begin{aligned} T(n+1|n) &= \overbrace{\left(\frac{n}{N}\right)\left(\frac{N-n}{N-1}\right)}^{\text{A born/B dies}} + \overbrace{\mu\left(\frac{N-n}{N}\right)}^{\text{B mutates to A}}, \\ T(n-1|n) &= \underbrace{\left(\frac{N-n}{N}\right)\left(\frac{n}{N-1}\right)}_{\text{B born/A dies}} + \underbrace{\mu\left(\frac{n}{N}\right)}_{\text{A mutates to B}} . \end{aligned} \tag{2.1.1}$$

Note that although these are written in terms of $T(n'|n)$, these can easily be amended to give $T(n|n')$; for instance $T(n|n+1) = T([n+1]-1|[n+1]) \equiv T(n-1|n)|_{n\to n+1}$. Also note that these transitions naturally respect the boundary conditions for the problem; $T(-1|0) = 0$ (you can't transition into a state with negative individuals) and $T(1|0) = \mu$ (if you start with no type A's, they can only enter the population through mutation).

We can now substitue Eqs. (2.1.1) into the master equation Eq. (2.0.1) to obtain an explicit expression for the dynamics of the probability of being in state n at time t;

$$\begin{aligned} \frac{dP_n(t)}{dt} = &\left[\frac{n+1}{N}\frac{(N-n-1)}{(N-1)}P_{n+1}(t) + \frac{n-1}{N}\frac{(N-n+1)}{(N-1)}P_{n-1}(t) - \left(2\frac{n}{N}\frac{(N-n)}{(N-1)}\right)P_n(t)\right] \\ &+ \mu\left[\frac{n+1}{N}P(n+1,t) + \frac{N-n+1}{N}P(n-1,t) - \left(\frac{n}{N}+\frac{N-n}{N}\right)P(n,t)\right]. \end{aligned} \tag{2.1.2}$$

While complete, this description is a little complicated. It can be simplified notationally by noting that the system is linear, such that for a birth-death process in 1D, we can write

$$\frac{d\boldsymbol{P}(t)}{dt} = M\boldsymbol{P}(t) \tag{2.1.3}$$

where $\boldsymbol{P}(t)$ is the $N+1$ length vector of probabilities $(P_0(t), P_1(t), \dots P_N(t))$, and M is a tridiagonal matrix. Various useful quantities can be derived from this equation; for instance the stationary distribution, which is independent of time, is given by $\frac{d\boldsymbol{P}^{\text{st}}}{dt} = \mathbf{0}$, or $M\boldsymbol{P}^{\text{st}} = \mathbf{0}$, a simple eigenvector problem. However deriving other quantities of interest becomes more complicated. We therefore seek an alternative approach in Section 2.2.

2.2. Derivation of the FPE and SDE from the master equation.

While simulating the master equation has been shown to be relatively straightforward, we are still no further in making the analytic progress which we initially sought. It will now be shown that the problem can be simplified significantly by using an approximation which resembles an approximation called the the Kramers-Moyal expansion. The Kramers-Moyal expansion is one of a set of schema which approximate the master equation by assuming a continuous state space. It has the same basic features of the *diffusion approximation* in population genetics. The van-Kampen system size expansion, or linear noise approximation (LNA), is another such scheme. If dealing with the scientific literature, note that these approximations essentially attempt the same thing; exploiting a large parameter and transitions between neighbouring states to approximate the state-space as continuous.

To illustrate the idea, the procedure will be applied to the specific example of the Moran model with mutation in Section 2.2.1, before a more general treatment is provided in Section 2.2.2.

2.2.1. *The two-allele haploid Moran model with mutation: FPE*

We begin by taking the master equation, Eq. (2.1.2), and introduce a new variable x such that $x = n/N$. The master equation then becomes

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} = &\Bigg[(x+\frac{1}{N})\frac{N}{N-1}(1-x-\frac{1}{N})p(x+1/N,t)\\ &+(x-\frac{1}{N})\frac{N}{N-1}(1-x+\frac{1}{N})p(x-1/N,t) - \left(2x\frac{N}{N-1}(1-x)\right)p(x,t)\Bigg]\\ &+\mu\left[\left(x+\frac{1}{N}\right)p(x+1/N,t)\right.\\ &\left.+\left(1-x+\frac{1}{N}\right)p(x-1/N,t) - (x+(1-x))\,p(x,t)\right]\,,\end{aligned} \tag{2.2.1}$$

where we note $p(x,t)$ is a new continuous distribution. The recurrent factors of $1/N$ in this equation give us a clue as how to proceed. If N is large, a Taylor expansion of $p(x,t)$ about x can be conducted. Assuming that the mutation rate is small (of order N^{-1}) and collecting terms order by order in $1/N$, one arrives at a one-dimensional example of the Fokker-Planck equation;

$$\frac{\partial p(x,t)}{\partial t} = -\frac{1}{N}\frac{\partial}{\partial x}\left[\mu(1-2x)p(x,t)\right] + \frac{1}{2N^2}\frac{\partial^2}{\partial x^2}\left[2x(1-x)p(x,t)\right] + \mathcal{O}(N^3) \tag{2.2.2}$$

Ignoring terms of order $\mathcal{O}(N^{-3})$, we arrive at an FPE (see Eq. 2.0.2). This PDE in two variables, x and t, is far more amenable to analysis than the $N+1$ partial difference equations comprising the master equation. Its physical interpretation is perhaps most evident when read as a convection-diffusion equation; the term preceding $p(x,t)$ in the first spatial derivative governs its 'bulk advection', while that in the second derivative describes diffusion. For this reason they are referred to as the drift (advection) and diffusion (noise) terms respectively[21]. We can also write a continuity equation for this system by introducing the probability flux or probability current, $J(x,t)$;

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x}\,. \tag{2.2.3}$$

Say we want to calculate how the mean value of x, $\langle x \rangle = \int xp(x,t)dx$, evolves in time. We can derive an equation for its time-evolution by multiplying Eq. (2.2.2) by x and integrating over all x. The expression can be further simplified by noting that $J(x,t)$ must be zero at the reflecting boundaries $x=0$ and $x=1$ and that in this particular model the diffusion term is also zero at the boundaries. Letting $\tau = t/N$, the resulting equation is then

$$\frac{d\langle x \rangle}{d\tau} = \mu(1-2x)\,. \tag{2.2.4}$$

The time-evolution of the mean is governed entirely by the advection term. This behaviour can also be obtained by rescaling time in Eq. (2.2.2) such that $\tau = t/N$, taking the limit $N \to \infty$ and noting that $p(x',t) = \delta(x(\tau),x')$ is a solution of the resulting equation. We will call this the macroscopic behaviour of the system, or alternatively the deterministic limit. Note that for a neutral model, with no mutation, there are no deterministic dynamics; thus in this limit the dynamics are *entirely* driven by noise.

A similar approach to that described above (which can be understood as accounting for variance as well as the mean) allows us to arrive at an SDE for the dynamics. We find

$$\frac{d\langle x \rangle}{d\tau} = \mu(1-2x) + \frac{1}{\sqrt{N}}\eta(x,\tau)\,, \tag{2.2.5}$$

with $\langle \eta(x,\tau)\eta(x,\tau')\rangle = 2x(1-x)$.

The approach described in this section allows us to move from a master equation to an analogous FPE or SDE. However this direct approach is rather cumbersome, and becomes difficult for more complicated models. Luckily a streamlined approach exists, which we explore in the following section.

[21] An unfortunate clash of nomenclature appears here between the physics and biology communities. In population genetics, genetic drift is the process by which the composition of a population is changed by noise. Drift in the context of population genetics therefore refers to the noisy component of a system's behaviour, rather than the deterministic component, as in physics.

2.2.2. *General multivariate FPE and SDE from master equation*

The expansion of the master equation is generalised as follows. Let us postulate that there is some large parameter N which is both inversely proportional to the reaction rates and some measure of the typical number of individuals in the population. In the case of the Moran model this is specifically the population size, though in general it could be the typical size (or volume) of the system which governs the interaction rate. A new set of variables $\boldsymbol{x} = \boldsymbol{n}/N$ is introduced. If N is a measure of population size, the new variables can be naturally interpreted as some measure of the concentration of each species in the population.

We begin by describing a system comprised of m different "species", where the word species is used to generically determine the different classes of individual that we are keeping track of in our probabilistic model (e.g. this could be the number of alleles, minus one due to the fixed population size constrain, in our moran model). The state of the system is given by an m-dimensional state vector $\boldsymbol{n}$. It is now useful to introduce the stoichiometric matrix, ν. The stoichiometric matrix is an m by u matrix which gives a concise way of stating which species were transformed in a given reaction. Each element ν_{ij} $(i = 1 \ldots m,\ j = 1 \ldots u)$ gives the change in number of the i^{th} species due to the j^{th} transition, or "reaction". The u reactions then take the system from state $\boldsymbol{n}$ to $\boldsymbol{n}' = \boldsymbol{n} + \boldsymbol{\nu}_j$, where the vector $\boldsymbol{\nu}_j$ is the j^{th} column of the matrix ν.

In this notation, for a given state $\boldsymbol{n}$ at time t, we can describe all the transitions by $T_\mu(\boldsymbol{n}+\boldsymbol{\nu}_\mu|\boldsymbol{n})$. The master equation may be rewritten in this new notation as

$$\frac{dP_{\boldsymbol{n}}, t}{dt} = \sum_{j=1}^{u} \left[T_j(\boldsymbol{n}|\boldsymbol{n}-\boldsymbol{\nu}_j)P_{\boldsymbol{n}-\boldsymbol{\nu}_j}(t) - T_j(\boldsymbol{n}+\boldsymbol{\nu}_j|\boldsymbol{n})P_{\boldsymbol{n}}(t) \right] . \tag{2.2.6}$$

In the Moran model with mutation, we simply have $m = 1$ and $u = 2$, while the stoichiometry matrix is

$$\nu = (1\,,\ -1) \tag{2.2.7}$$

since the system is described by a single species variable and the transitions either increase or decrease this number by one.

We now make a continuous approximation of the transition rates. Introducing the functions $f_j(\boldsymbol{x}) = T_j(N\boldsymbol{x} + \boldsymbol{\nu}_j|N\boldsymbol{x})|_{N\to\infty}$, the master equation 2.2.6 can be reexpressed

$$\frac{dp(\boldsymbol{x},t)}{dt} = \sum_{j=1}^{u} \left[f_j(\boldsymbol{x}-\boldsymbol{\nu}_j N^{-1})p(\boldsymbol{x}-\boldsymbol{\nu}_j N^{-1},t) - f_j(\boldsymbol{x})p(\boldsymbol{x},t) \right] , \tag{2.2.8}$$

where we have again changed from a distribution $P(\boldsymbol{n},t)$ to $p(\boldsymbol{x},t)$. In this form it is clear that if N is large, one may proceed in the same way as in the Moran model example by assuming $\boldsymbol{x}$ is continuous and implementing a Taylor expansion about $\boldsymbol{x}$. This is the essence of the master equation expansion. For the Moran model with mutation we have

$$\begin{aligned}
f_1(x) &= T(Nx+1|Nx)\\
&= \left[\left(\frac{xN}{N}\right)\left(\frac{N-xN}{N-1}\right)+\mu\left(\frac{N-xN}{N}\right)\right]\Bigg|_{N\to\infty}\\
&= \left[x(1-x)\left(\frac{N}{N-1}\right)+\mu(1-x)\right]\Bigg|_{N\to\infty}\\
&= x(1-x)+\mu(1-x)\\
f_2(x) &= T(Nx-1|Nx)\\
&= \left[\left(\frac{N-xN}{N}\right)\left(\frac{xN}{N-1}\right)+\mu\left(\frac{xN}{N}\right)\right]\Bigg|_{N\to\infty}\\
&= \left[x(1-x)\left(\frac{N}{N-1}\right)+\mu x\right]\Bigg|_{N\to\infty}\\
&= x(1-x)+\mu x\,.
\end{aligned} \tag{2.2.9}$$

Notice that these $f_j(x)$ terms should be independent of N.

Conducting a Taylor expansion of Eq. (2.2.8) and truncating after order N^{-2}, we obtain a multivariate FPE which takes the form

$$\frac{\partial p(\boldsymbol{x},t)}{\partial t} = -\frac{1}{N}\sum_{i=1}^{m}\frac{\partial}{\partial x_i}\left[A_i(\boldsymbol{x})p(\boldsymbol{x},t)\right]+\frac{1}{2N^2}\sum_{i.j=1}^{m}\frac{\partial^2}{\partial x_i\partial x_j}\left[B_{ij}(\boldsymbol{x})p(\boldsymbol{x},t)\right]\,, \tag{2.2.10}$$

where $\boldsymbol{A}(\boldsymbol{x})$ is is now the drift/advection vector, while $B(\boldsymbol{x})$ is the diffusion/noise matrix. Their forms are governed by the stoichiometry matrix and the transition rates [2] such that

$$A_i(\boldsymbol{x}) = \lim_{N\to\infty}\sum_{\mu=1}^{u}\nu_{i\mu}f_\mu(\boldsymbol{x}) \tag{2.2.11}$$

and

$$B_{ij}(\boldsymbol{x}) = \lim_{N\to\infty}\sum_{\mu=1}^{u}\nu_{i\mu}\nu_{j\mu}f_\mu(\boldsymbol{x})\,. \tag{2.2.12}$$

Once again, rescaling time such that $\tau = t/N$ and taking the limit $N\to\infty$, the resulting equation admits $P(\boldsymbol{x}',t)=\delta(\boldsymbol{x}(\tau),\boldsymbol{x}')$ as a solution such that

$$\frac{d\boldsymbol{x}}{d\tau} = \boldsymbol{A}(\boldsymbol{x})\,, \tag{2.2.13}$$

describes the deterministic, macroscopic dynamics. Similarly the analogous multivariate SDE can be shown to take the form

$$\frac{d\boldsymbol{x}}{d\tau} = \boldsymbol{A}(\boldsymbol{x})+\frac{1}{\sqrt{N}}\boldsymbol{\eta}(\boldsymbol{x},t)\,, \tag{2.2.14}$$

with $\langle\eta_i(\boldsymbol{x},\tau)\eta_j(\boldsymbol{x},\tau')\rangle = B_{ij}(\boldsymbol{x})\delta\left(\tau-\tau'\right)$

Since it is possible to calculate $\boldsymbol{A}(\boldsymbol{x})$ and $B(\boldsymbol{x})$ entirely from the stoichiometry matrix and reaction rates, models are sometimes expressed in the notation of chemical reactions, in terms of reactants and products (in fact the term 'stoichiometric matrix' is borrowed from chemistry). For an arbitrary m-dimensional IBM, whose dynamics are fully described by a set of u reaction rates, the model can be expressed in chemical reaction notation as

$$\sum_{i=1}^{m} a_{ji} X_i \xrightarrow{r_j} \sum_{i=1}^{m} b_{ji} X_i, \quad \forall j = 1, \dots u, \tag{2.2.15}$$

where a_{ji} and b_{ji} respectively specify the reactants and products of the j^{th} reaction, and r_j are the reaction rate constants. The elements of the stoichiometric matrix are then given by $\nu_{ij} = b_{ji} - a_{ji}$, while the reaction rate constants r_j are related to the transition rates by an equation of the form

$$T_j(\boldsymbol{n}'|\boldsymbol{n}) \propto r_j \prod_{i=1}^{m} a_{ji} \frac{n_i}{N}\,. \tag{2.2.16}$$

Implicit in this notation is the assumption that the probability of a reaction occurring is proportional to the product of the reactant concentrations. This is also known as the law of mass action. Most often this is the case, however situations may arise in which we wish to incorporate additional state dependence (e.g. a Hollings type II uptake rate).

In expanding the master equation, we have moved from the microscopic description of a model involving as many equations as there are states (the master equation) to one in which there are only as many variables as there are species (the FPE). The crux of the approximation is in changing to a new set of variables $\boldsymbol{x} = \boldsymbol{n}/N$ which are approximately continuous (the diffusion approximation) before applying a Taylor expansion to the master equation and neglecting terms of order N^{-3}. When modelling populations, the parameter N is usually identified as the size of the population. In the limit $N \to \infty$, we have seen that the FPE describes a deterministic dynamic which we have termed the macroscopic dynamic. In this spirit, since the FPE lies between the master equation and deterministic equation in terms of detail, it is often referred to as a mesoscopic description.

The Fokker-Planck equation is clearly more amenable to analysis than the master equation with which we began. Of particular interest is the one-dimensional Fokker-Planck equation, from which many properties of interest can be calculated analytically, as we explore in the following section.

2.3. The FPE: Some useful manipulations.

We have already seen that the FPE can be seen as in some sense analogous to an SDE. This is useful in an of itself; it tells us that as $N \to \infty$ (i.e. in infinitely large populations) the dynamics of the stochastic system are in fact no longer stochastic, but given by deterministic dynamics governed by $\boldsymbol{A}(\boldsymbol{x})$. For very large populations, we can therefore imagine trajectories that appear almost deterministic, but with small fluctuations around the deterministic behaviour. However this does not specify what constitutes a "very large population". We will see that this can depend on the other parameters in the problem. To elucidate this, we will look at some of the stochastic characteristics obtainable from the FPE. These include the stationary probability distribution, fixation probability, and fixation time.

We begin by imagining the system existing on an interval $[a_1, a_2]$ in state space, from which it cannot leave. The FPE (2.2.2) describes such a system, with x lying on the interval $[0, 1]$. In this case there should be zero flow of probability across the boundary. Therefore the probability current must be zero when evaluated at the boundaries; $J(a_1, t) = J(a_2, t) = 0$. Such boundaries are called reflecting.

Now let us consider a system in which there exist states where there are no dynamics. Such states are called absorbing. In order to account for this, we define the barriers as existing outside of the interval, so that when the system reaches the boundary it is removed. The probability of being at either of the boundaries is then zero, $p(a_1, t) = p(a_2, t) = 0$. If the barriers a_1 and a_2 of the system are absorbing, it is clear that given an infinite amount of time, the probability of the system remaining on the interval will tend to zero as probability 'leaks out' of the interval. An example of such a system is the neutral Moran model with we've been looking at with mutation switched off ($\mu = 0$). Finally, if the boundaries are at infinity, we expect $p(x, t)$, along with $J(x, t)$, to vanish at $x = \pm\infty$.

2.3.1. *The stationary probability distribution*

First consider a system with two reflecting boundaries at a_1 and a_2. In the limit of long times one might expect the PDF $p(x, t)$ in Eq. (2.0.2) to become independent of time. We call this distribution the stationary distribution, and it is defined by

$$p_{\mathrm{st}}(x) = \lim_{t\to\infty} p(x, t)\,. \tag{2.3.1}$$

Since $p_{\mathrm{st}}(x)$ is independent of time, we find it must satisfy the equation

$$-\frac{1}{N}\frac{d}{dx}\left[A(x)p_{\mathrm{st}}(x)\right] + \frac{1}{2N^2}\frac{d^2}{dx^2}\left[B(x)p_{\mathrm{st}}(x)\right] = 0\,. \tag{2.3.2}$$

The solution to this equation is

$$p_{\mathrm{st}}(x) = \exp\left(\int_{a_1}^{x} dy \frac{2NA(y) - dB(y)/dy}{B(y)}\right)\left[\int_{a_1}^{a_2} dx \exp\left(\int_{a_1}^{x} dy \frac{2NA(y) - dB(y)/dy}{B(y)}\right)\right]^{-1}\,,$$

or

$$p_{\mathrm{st}}(x) = \frac{1}{B(x)}\exp\left(2N\int_{a_1}^{x} dy \frac{A(y)}{B(y)}\right)\left[\int_{a_1}^{a_2} dx \frac{1}{B(x)}\exp\left(2N\int_{a_1}^{x} dy \frac{A(y)}{B(y)}\right)\right]^{-1} \tag{2.3.3}$$

where we have appropriately normalised so that $\int_{a_1}^{a_2} dx\, p_{\mathrm{st}}(x) = 1$. This can yield interesting results about the system's long time behaviour. For instance, we can

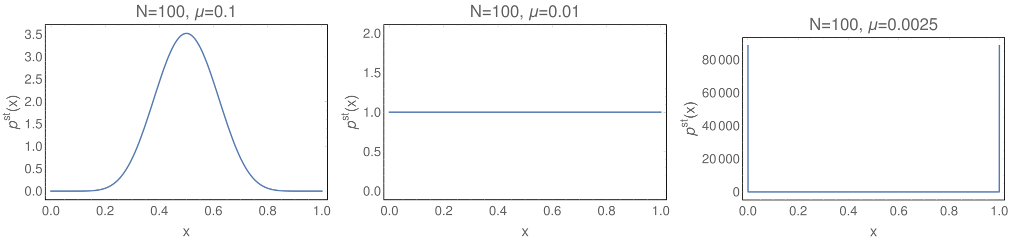

FIGURE 1. Stationary probability distributions obtained from Eq. (2.3.5) for various μ. When μN is large (left panel) the distribution is centered on the model's deterministic fixed point, with noise (genetic drift) generated by the finite population size driving some variance around this. When μN is small (right panel) the distribution converges to two delta peaks at $x = 0$ and $x = 1$; mutations at a per-capita rate μ are entering the population slowly enough that mutations are driven to extinction or fixation on a much faster timescale than the arrival rate of new mutations. The system therefore spends a lot of time at $x = 0$ or $x = 1$ waiting for a new mutation to appear, so that over long times we see an equal probability of being in either state. In the middle plot we see that the transition between these deterministic and stochastic regimes occurs when $\mu N = 1$.

look at the FPE for the Moran model with mutation, Eq. (2.2.2), in which there are no absorbing states for $\mu > 0$. For generality, we write

$$A(x) = \mu(1 - 2x)\,, \qquad B(x) = 2x(1 - x)\,. \tag{2.3.4}$$

Substituting these terms into Eq. (2.3.3), and noting that x lies on the interval $[0, 1]$, the stationary distribution takes the form

$$p_{\mathrm{st}}(x) = \frac{x^c(1-x)^c}{\int_0^1 dx x^c(1-x)^c}\,, \tag{2.3.5}$$

with

$$c = N\mu - 1\,.$$

Some of the different behaviour the system can exhibit is demonstrated in Fig. 2.

If the system instead has absorbing boundaries, there is no stationary distribution on the interval $a_1 < x < a_2$, since all the stochastic trajectories eventually leave the interval. In such cases, there are other measures of the system's behaviour which are more illuminating.

2.3.2. *First passage problems*

First passage problems are those which ask 'what is the probability the system reaches a particular final condition?', or 'what are the statistics of the time for the system to reach this final condition?'. In order to calculate these first passage properties, we find it useful to work with the backward Fokker-Planck equation (BFPE). This can be expressed in one dimension as

$$-\frac{\partial q(x,t|x_0,t_0)}{\partial t_0} = \frac{A(x_0)}{N}\frac{\partial}{\partial x_0}\left[q(x,t|x_0,t_0)\right] + \frac{B(x_0)}{2N^2}\frac{\partial^2}{\partial x_0^2}\left[q(x,t|x_0,t_0)\right] . \quad (2.3.6)$$

If the process under consideration is homogeneous (time-independent), the evolution of the system depends only on the difference between the initial and final time $t-t_0$, and the BFPE can be rewritten in terms of the derivative with respect to t;

$$\frac{\partial q(x,t|x_0,t_0)}{\partial t} = \frac{A(x_0)}{N}\frac{\partial}{\partial x_0}\left[q(x,t|x_0,t_0)\right] + \frac{B(x_0)}{2N^2}\frac{\partial^2}{\partial x_0^2}\left[q(x,t|x_0,t_0)\right] . \quad (2.3.7)$$

The key difference between the forward FPE (2.2.10) and the BFPE, is which set of variables are kept fixed and which vary. In the forward FPE the initial conditions x_0 at t_0 are kept fixed, and one finds for solutions for $t \geq t_0$. In the BFPE we keep the final condition x at t fixed and calculate for solutions with $t_0 \leq t$. Since in the BFPE we fix the final condition, it is clearly more useful when dealing with first passage problems. For simplicity, we restrict ourselves to a one-dimensional homogeneous system.

2.3.3. *Unconditional Fixation Time*

We wish to know the time until a system first escapes the region between two points, $x = a_1$ and $x = a_2$, given some initial condition $a_2 > x_0 > a_1$. This time is clearly a stochastic variable and so it will be described by a PDF indicating the probability of a certain first passage time t given initial condition x_0. It is denoted here by $\mathcal{T}(x_0,t)$. We begin by defining the probability $G(x_0,t)$ that, at some time t, the system is still on the interval;

$$G(x_0,t) = \int_{a_1}^{a_2} dx\, q(x,t|x_0,t_0)\,, \quad (2.3.8)$$

where the dependence on the initial time has been suppressed. Integrating Eq. (2.3.7) over x between a_1 and a_2, we find in fact the equation for $G(x_0,t)$ obeys the same BFPE as $q(x,t|x_0,0)$;

$$\frac{\partial G(x_0,t)}{\partial t} = \frac{A(x_0)}{N}\frac{\partial}{\partial x_0}\left[G(x_0,t)\right] + \frac{B(x_0)}{2N^2}\frac{\partial^2}{\partial x_0^2}\left[G(x_0,t)\right]\,, \quad (2.3.9)$$

with initial condition

$$G(x_0,t_0) = 1 \quad \text{if } a_1 < x_0 < a_2\,, \quad (2.3.10)$$

$$G(x_0,t_0) = 0 \quad \text{elsewhere}\,, \quad (2.3.11)$$

and boundary conditions

$$G(a_1,t) = G(a_2,t) = 0\,. \quad (2.3.12)$$

Since $G(x_0,t)$ is the probability that at time t the system is still on the interval $a_1 < x < a_2$, the quantity $G(x_0,t) - G(x_0,t+\Delta t)$ is the probability that the system has reached one of the boundaries during t to $t+\Delta t$. This can be related to $\mathcal{T}(x_0,t)$ quite simply, since

$$\mathcal{T}(x_0,t)\Delta t = G(x_0,t) - G(x_0,t+\Delta t)\,, \tag{2.3.13}$$

or, rearranging and sending $\Delta t \to 0$,

$$\mathcal{T}(x_0,t) = -\frac{\partial G(x_0,t)}{\partial t}\,. \tag{2.3.14}$$

The mean time for the system to leave the interval, denoted $T(x_0)$, can be calculated directly from the distribution $\mathcal{T}(x_0,t)$ by $T(x_0) = \int_{t_0}^{\infty} t\mathcal{T}(x_0,t)\,dt$. Using the above equality, this can be expressed

$$T(x_0) = -\int_{t_0}^{\infty} t\frac{\partial G(x_0,t)}{\partial t}dt\,. \tag{2.3.15}$$

This equation can be integrated by parts; letting $t_0 = 0$ and assuming that $tG(x_0,t)$ tends to zero as $t \to \infty$, one arrives at

$$T(x_0) = \int_{0}^{\infty} G(x_0,t)dt\,. \tag{2.3.16}$$

Integrating Eq. (2.3.9) over time and noting that $G(x_0,t)$ tends to zero as $t \to \infty$, we then arrive at an equation for the mean time to reach either of the boundaries

$$-1 = \frac{A(x_0)}{N}\frac{d}{dx_0}T(x_0) + \frac{B(x_0)}{2N^2}\frac{d^2}{dx_0^2}T(x_0) \tag{2.3.17}$$

with the boundary conditions,

$$T(0) = 0\,, \qquad T(1) = 0\,. \tag{2.3.18}$$

In this case the boundary conditions follow by noting that they are the time to reach *either* $x = a_1$ or $x = a_2$, so that at both extremes the system has already fixed on the boundaries.

Let us go back to the neutral Moran model described by Eq. (2.3.4), but with $\mu = 0$ (i.e. switching mutation off). In this case the points $x = 0$ and $x = 1$ are absorbing boundaries from which the system cannot leave. At these points the population is said to have fixated. Using Eq.(2.3.17) and Eq. (2.3.18), with $A(x) = 0$ and $B(x)$ taken from Eq. (2.3.4), we can calculate the time this would take to happen. One finds

$$T(x_0) \;\; = \;\; -N^2\left[(1-x_0)\ln{(1-x_0)} + x_0\ln{(x_0)}\right]\,, \tag{2.3.19}$$

which is plotted in the right panel of Figure 2. In the nomenclature of population genetics this is called the mean unconditional fixation time.

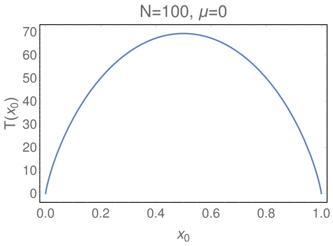

FIGURE 2. Unconditional fixation time for the neutral moran model, taken from Eq. (2.3.19). The time for either A or B to go extinct is intuitively maximised when $x_0 = 1/2$. At this point the fixation time is $-N\ln(2)$ when measured in time units τ, or $-N\ln(2)$ in time units of t. Thus the extinction time scales like N^2 in the natural units of the birth-death process (where the average time between birth-death events is 1).

2.3.4. *Fixation Probability*

What about the probability that the system hits one of these boundaries before the other? Let us return to considering a function of the same form as $G(x_0, t)$ in Eq. (2.3.8), but this time introduce two slightly different functions to account for different integration limits;

$$G_{a_1}(x_0, t) = 1 - \int_{a_1}^{\infty} dx\, q(x, t|x_0, t_0) = \int_{-\infty}^{a_1} dx\, q(x, t|x_0, t_0)\,, \qquad (2.3.20)$$

$$G_{a_2}(x_0, t) = \int_{a_2}^{\infty} dx\, q(x, t|x_0, t_0) = 1 - \int_{-\infty}^{a_2} dx\, q(x, t|x_0, t_0)\,. \qquad (2.3.21)$$

The first of these functions gives the probability that, at time t, the system is at some point $x < a_1$ and the second that it is at some point $x > a_2$. Of course, this gives us no information about which of these regions the system ended up in *first*. To do this would require a consideration of the trajectories conditioned such that the time to one boundary was less that the time to the other. Our task is significantly simplified however if we force the boundaries a_1 and a_2 to be absorbing. Then once the system hits state a_1 or a_2 it is immediately removed and we do not have to worry about time-ordering. The functions $G_{a_1}(x_0, t)$ and $G_{a_2}(x_0, t)$ then tell us respectively whether at time t the system has hit either a_1 or a_2. Both functions obey the equations

$$\frac{\partial G_{a_1/a_2}(x_0, t)}{\partial t_0} = \frac{A(x_0)}{N}\frac{\partial}{\partial x_0}\left[G_{a_1/a_2}(x_0, t)\right] + \frac{B(x_0)}{2N^2}\frac{\partial^2}{\partial x_0^2}\left[G_{a_1/a_2}(x_0, t)\right] \quad (2.3.22)$$

but with different boundary conditions. As $t \to \infty$, the probability of having hit either a_1 or a_2 tends to one. Introducing

$$Q_{a_1/a_2}(x_0) = \lim_{t\to\infty} G_{a_1/a_2}(x_0, t)\,, \qquad (2.3.23)$$

we see the equation for both functions is

$$0 = \frac{A(x_0)}{N}\frac{\partial}{\partial x_0}\left[Q_{a_1/a_2}(x_0)\right] + \frac{B(x_0)}{2N^2}\frac{\partial^2}{\partial x_0^2}\left[Q_{a_1/a_2}(x_0)\right] \,, \tag{2.3.24}$$

albeit with different boundary conditions. For $Q_{a_1}(x_0)$ we have

$$Q_{a_1}(a_1) = 1\,, \qquad Q_{a_1}(a_2) = 0\,, \tag{2.3.25}$$

and for $Q_{a_2}(x_0)$ instead

$$Q_{a_2}(a_1) = 0\,, \qquad Q_{a_2}(a_2) = 1\,. \tag{2.3.26}$$

Once again the neutral Moran model described by Eq. (2.3.4), but with $\mu = 0$ (i.e. switching mutation off), may be used to illustrate the method. We ask the question, what is the probability of the system reaching the point $x = 1$ given some initial condition x_0? Since at $x = 1$ the system is composed entirely of the A type individuals, this is called the fixation probability. In the neutral case $A(x_0) = 0$ and therefore we obtain

$$Q_1(x_0) = x_0\,, \qquad Q_0(x_0) = 1 - x_0\,. \tag{2.3.27}$$

The probability of either type fixating is thus simply proportional to their respective initial frequencies in the neutral model. We note that in population genetics it is most common to simply discuss the probability of fixation of the A type and for simplicity we will often write this probability $Q(x_0) \equiv Q_1(x_0)$.

2.3.5. *Summary*

In deterministic systems, we are interested in the stability of fixed points; at long times these systems converge to fixed points that are stable. However in stochastic systems, such convergence is not guaranteed. Instead events can send a population to extinction. With this in mind, when analysing the stochastic system, we ask a different suite of questions; these include *"what is the probability of extinction (or fixation)?"*, *"how long does this extinction take?"*, and *"what is the probability of finding the system in some state at long times?"*. These questions can be answered by determining the fixation probability, fixation time, and stationary probability distribution respectively.

Although the derivations in Sections 2.3.1-2.3.4 get quite technical, the important thing to note is that in 1D Eqs. (2.3.3), (2.3.17), and (2.3.24) provide off-the-shelf equations with which to solve for the stationary distribution, unconditional fixation time, and fixation probability of any one-dimensional problem of interest. Of course, this restricts us to 1D problems. Problems in higher dimensions are significantly more difficult to handle. A common approach therefore when dealing with multivariate systems is to reduce to the dimensionality of the problem. Fast-variable elimination is one possible route, as we have learned at the start of this module. However one must take care when using this approach in stochastic systems.