Diffusion Process Models in Mathematical Genetics

Alison Etheridge

Warning: These notes have been typed in great haste and may contain inaccuracies. They should not be regarded as a substitute for going to lectures.

Background and contents

In this half of this term's course we are going to take a very different approach to mathematical genetics. To a large extent so far you have been looking backwards in time – constructing genealogical trees relating individuals in a sample and using these as a route to understanding the partitions that one should expect to observe in genetic data. Now we're going to be concentrating far more on the corresponding *forwards* in time models which describe how the population from which you are drawing your sample evolves. You began term with one such model, the Wright-Fisher model, but that is very special and we shall instead develop a model -valid for large populations - that approximates the Wright-Fisher model, but also other classical population genetics models. This model is an example of what is known as a *diffusion approximation* and for quite a bit of our time we'll be trying to gain a little understanding of what it means to be a diffusion. Reassuringly, we'll see that our diffusion approximation fits precisely with the coalescent approximation that you have been studying so far.

The diffusion approach to population genetics is very classical, certainly much older than the coalescent and although for a while it looked as though it would be eclipsed, more recently it has become clear that the key to understanding some of the more biologically realistic models is an amalgam of the two approaches. The diffusion approach will provide us with a whole new set of techniques.

Here then is an outline of the rest of the course.

1. **The Moran model.** We'll begin by introducing another classical model of population genetics. Although less popular than the Wright-Fisher model among biologists, we'll see that in fact it retains key features of the Wright-Fisher model and indeed for large populations the two models can be regarded as being 'close' to one another. Mathematically, the Moran model is a *birth and death* process which makes it analytically much more tractable.

2. **The infinite population limit.** In this short section we describe briefly how to approximate the Moran model by passing to an infinite population limit. We'll describe (but not provide a rigorous justification for) a beautiful construction of the infinite limit due to Peter Donnelly and Tom Kurtz which retains the genealogy of the population, thus showing that our infinite population limit really does correspond to the coalescent model that you have been studying.

3. **Diffusions.** The limit that we obtain is what is known as a diffusion process and in this section we step back from the genetics to take a look at what a diffusion process is and how it can be characterised. We briefly mention stochastic differential equations, but only as an heuristic representation. For the most part we shall be concerned with what is known as the generator of the process.

4. **Speed and Scale.** In one dimension all diffusions can be obtained from a special one called Brownian motion by certain transformations of space and time. We exploit this to find formulae for some quantities of genetic interest.

5. **Selection and Mutation.** Everything so far has been concerned with the most basic model. In this section we increase the biological sophistification somewhat. Here for the first time we reap the benefits of the diffusion approximation. We obtain exact expressions for quantities in the diffusion world where in the Wright-Fisher and Moran models this was either impossible or so complicated as to obscure the real effects of the different genetic processes.

6. **More than two types: Dirichlet and Poisson-Dirichlet distributions.** So far we have had only two alleles in our population. Now we extend this to mutliple allelic types and in an important special case uncover the stationary distribution of the allele frequency as being governed by the Dirichlet distribution or in the infinitely many alleles limit the Poisson-Dirichlet distribution. We'll be able to answer questions like 'What is the probability that an allele that is at frequency x in the population is in fact the oldest?'

7. **Ewens' Sampling Formula revisited.** Finally (time permitting) we give a very simple derivation of the Ewens Sampling Formula.

1 The Moran model

First let's remind ourselves about the basic (forwards in time) model for the evolution of our population that was introduced in Bob Griffiths's first lecture.

Definition 1.1 (The neutral Wright-Fisher model) *The neutral Wright-Fisher model is described as follows. A population of N genes evolves in discrete generations. Generation $(k+1)$ is formed from generation k by choosing N genes at random with replacement. i.e. each gene in generation $(k+1)$ chooses its parent at random from those present in generation k.*

In this model some genes have no offspring, others may have several.

From this definition it is an elementary matter to work out the genealogical trees that relate individuals in a sample from the population. To remind you how this worked, suppose first that we take a sample of size two from the population. The probability that these two individuals share a common parent in the previous generation is $\frac{1}{N}$. If they do not, then the probability that their parents had a common parent is $\frac{1}{N}$, and so on. In other words, the time to the most recent common ancestor (MRCA) of the two individuals in the sample has a geometric distribution with success probability $\frac{1}{N}$. (The probability that their MRCA was k generations in the past is pq^{k-1} where $p = \frac{1}{N}$ and $q = 1 - p$.) In particular, the expected number of generations back to their MRCA is N. Now typically we are interested in large populations, where our rather crude models have some hope of having something meaningful to say. Then it makes sense to measure time in units of size N and in those units the time to the MRCA of a sample of size two is approximately exponentially distributed with parameter one. More generally, for a sample of size k, since the probability of three (or more) individuals sharing a common parent is $\mathcal{O}\left(\frac{1}{N^2}\right)$ and similarly the chance that two separate pairs of individuals are 'siblings' is $\mathcal{O}\left(\frac{1}{N^2}\right)$, the time (in units of size N) before the present at which two individuals in our sample share a common ancestor is approximately exponentially distributed with rate $\binom{k}{2}$ and when that 'merger' in the ancestral lineages of the sample takes place, it is equally likely to be any of the $\binom{k}{2}$ pairs that merges. Another way to say this is that each of the $\binom{k}{2}$ pairs of individuals has an exponential random variable with parameter one associated with it. We think of these random variables as alarm clocks.

The first event in our genealogical tree as we trace backwards in time is the merger of the two lines whose alarm clock goes off first. (The minimum of $\binom{k}{2}$ exponential one random variables is exponential with paramater $\binom{k}{2}$.) After that we just trace the remaining $\binom{k-1}{2}$ pairs of lineages and the same picture holds. This is just a way of describing the Kingman coalescent which you already know so much about.

The Moran model which we now introduce will have the property that a sample from the population will still be related by Kingman's coalescent, but the forwards in time population model will be much simpler than the Wright-Fisher model to study.

There are two essential differences between the Wright-Fisher model and the Moran model. First, whereas the Wright-Fisher model evolves in discrete generations, in the Moran model generations overlap. Second, in the Wright-Fisher model an individual can have up to N offspring, but in the Moran model an individual always has zero or two offspring.

Definition 1.2 (The neutral Moran model) *A population of N genes (labelled $1, \ldots, N$) evolves according to the Moran model if at exponential rate $\binom{N}{2}$ a pair of genes is sampled (with replacement) from the population, one dies and the other splits in two.*

Remark 1.3 *There are many different parametrisations (that is choices of the exponential rate) to choose from. We have chosen a convenient one but there is no standard choice.*

What do we mean by exponential rate? Just that we wait for an exponentially distributed time and then a pair is picked. After the reproduction event the process goes on to evolve (independently) in the same way.

Equivalent to this is to say that our population is labelled $1, \ldots, N$. Each pair of labels has an alarm clock which will go off at intervals which are independently exponentially distributed with parameter one (we call this sequence of times a *Poisson process*) and when a clock goes off - corresponding to labels (i, j) say, one gene dies and the other reproduces (with equal probabilities). The offspring adopt the labels (i, j).

Graphically:

[PICTURE]

where we have drawn an arrow between the lines labelled (i, j) whenever the (i, j) clock rings. The arrow $i \to j$ indicates that i reproduced and j dies, $i \leftarrow j$ indicates that j reproduced and i died.

We can recover the ancestry of a sample by tracing backwards in time. If an ancestral line is at the tip of an arrow, then it *coalesces* with that at the root. If it is at the root is will be unaffected.

[PICTURE]

It is not hard to convince oneself that the genealogical trees from a sample are then precisely those generated by Kingman's coalescent.

For example, follow a sample of size two backwards in time. The labels of the two individuals will change with time, let's call them $(i(t), j(t))$ say, but because of the lack of memory property of the exponential distribution, the time until the clock corresponding to $(i(t), j(t))$ rings is still going to be exponential parameter one. [If an exponential random variable has not rung by time t, then the extra time we must wait until it rings is still exponential with the same parameter. This means that if, for example, $i(t)$ changes to $i(t+)$, the exponentials for pairs involving $i(t)$ can be pieced together with those for $i(t+)$ to produce exponentials again.]

This then tells us that for large populations the genealogy of a sample from the Wright-Fisher model has approximately the same distribution as that of a sample from the Moran model, at least provided that we measure time for the Wright-Fisher model in units of population size. (Notice that the Moran model already evolves in the coalescent timescale.)

So far, although we have implicitly assumed that all the genes in our population are selectively neutral - so that all have an equal chance of reproductive success - we have not associated *types* with our genes.

Suppose then that the gene in question has two alleles, labelled a and A say. A basic question is 'How do the frequencies of the two alleles evolve with time?'

For both our models, what happens next - the formation of the next generation in the Wright-Fisher model or the next reproductive event in the Moran model - depends only on the current frequency of types, not on the past history of the population. In other words, the frequency of types evolves according to a Markov chain. For the Wright-Fisher model it is a discrete time chain of the sort that you encountered in section A probability. The Moran model is a continuous time chain - which just means that the times between events are determined by a sequence of independent exponentially distributed random variables. But it is an especially simple Markov chain - known as a birth and death process - in which changes of numbers of alleles are only of size one at each event. In fact you have already seen a continuous time Markov chain in Bob's lectures - the number of ancestors of the present day sample alive at time t in the past is a pure death process.

We shall use the names Wright-Fisher model *and* Moran model *both for the full models described above and for the corresponding Markov chains which keep track just of the frequencies of the different alleles in the population.*

Let's tell/remind ourselves of a few basic facts about Markov chains. If you didn't do section A probability then you may find the lecture notes on the web useful.

Aside on Markov chains.

A stochastic process is just a model for a random quantity that evolves with time. In other words X is a collection of random variables $\{X_t : t \in T\}$ indexed by a set T which we'll interpret as time. For us T will always be either $\{0, 1, 2, \ldots\}$ or $[0, \infty)$ (discrete time and continuous time stochastic processes respectively).

We shall use the notation $\mathcal{F}_t$ to mean the history of the process up until time t - so the information available to us if we watched the process up to time t.

A stochastic process has the *Markov property* if its future evolution conditional on knowing all of $\mathcal{F}_t$ is the same as if we condition on knowing just X_t. In other words, where it goes next may depend on its current value, but not on how it got to that value.

This is most easily formalised if the random variables X_t are themselves discrete random variables - that is take their values in a finite or countable state space Ω. (For our Wright-Fisher and Moran models the random variable X_t will tell us how many a (or A) alleles are present at time t and the state space in both cases is just the finite set $\{0, 1, \ldots, N\}$.)

Definition 1.4 (Discrete time Markov chain) *Let $X = \{X_0, X_1, \ldots\}$ be a collection of random variables which take values in some countable set Ω. Then X is said to be a* discrete time Markov chain *if it satisfies the* Markov property*:*

$$\mathbb{P}[X_n = x_n | \mathcal{F}_{n-1}] = \mathbb{P}[X_n = x_n | X_{n-1} = x_{n-1}].$$

That is,

$$\mathbb{P}[X_n = x_n | X_0 = x_0, X_1 = x_1, \ldots, X_{n-1} = x_{n-1}] = \mathbb{P}[X_n = x_n | X_{n-1} = x_{n-1}],$$
$$\text{for all } x_0, x_1, \ldots, x_{n-1} \in \Omega.$$

For a continuous time stochastic process the Markov property becomes (still assuming that Ω is finite or countable)

$$\mathbb{P}[X_t = x | \mathcal{F}_s] = \mathbb{P}[X_t = x | X_s], \qquad \forall s < t, x \in \Omega.$$

A continuous time Markov chain is the same as a discrete time Markov chain except that the time between transitions is exponentially distributed. The 'lack of memory' of the exponential distribution guarantees that the (continuous time version of the) Markov property holds for a continuous time Markov chain.

For the examples that we care about, Ω will be a subset of the non-negative integers. We then write

$$p_{ij} = \mathbb{P}[X_{n+1} = j | X_n = i].$$

Example 1.5 *In our neutral Wright-Fisher model,*

$$p_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(\frac{N-i}{N}\right)^{N-j}, \quad \textit{for } j = 0, 1, \ldots, N.$$

To see this, recall that to form generation $n+1$, each gene chooses its parent at random (with replacement) from the i type a genes and the $N-i$ type A genes in the nth generation.

For a continuous time Markov chain one often studies the embedded discrete time chain which is obtained by replacing the times between events by discrete times.

Example 1.6 *For the Moran model, the embedded chain has transitions*

$$p_{ij} = \begin{cases} \frac{i(N-i)}{N^2} & \textit{if } j = i \pm 1 \\ \frac{i^2}{N^2} + \frac{(N-i)^2}{N^2} & j = i \\ 0 & \textit{otherwise.} \end{cases}$$

For both the Wright-Fisher and Moran models that we have described so far, once the process hits 0 or N it stays there. That is 0 and N are *absorbing states*.

One is often interested in the time until one of these states is hit and the probability of hitting zero before N.

Let us write $H^A = \inf\{n \geq 0 : X_n \in A\}$, ($\inf \emptyset = \infty$) for the *hitting time* of the set A by the chain and we'll write $k_i^A = \mathbb{E}[H^A | X_0 = i]$ for the expected value of this quantity if we start from i at time zero.

Theorem 1.7 *The vector of mean hitting times $k^A = (k_i^A : i \in \Omega)$ is the minimal non-negative solution to the system of linear equations*

$$k_i^A = 0, \qquad \textit{for } i \in A \tag{1}$$

$$k_i^A = 1 + \sum_{j \in \Omega} p_{ij} k_j^A, \qquad \textit{for } i \notin A. \tag{2}$$

A heuristic justification can be obtained by conditioning on the outcome of the first jump of the chain. (Minimality means that if $x = (x_i : i \in \Omega)$ is another solution with $x_i \geq 0$ for all i, then $x_i \geq k_i$ for all i.)

Example 1.8 *For the neutral Wright-Fisher model, 0 and N are absorbing states for the proportion of a alleles. For large populations (that is large N) if the initial frequency of a-alleles is p, then the expected time to absorption, is approximately $t(p)$ given by*

$$t(p) = -2N \left(p \log p + (1-p) \log(1-p)\right). \tag{3}$$

There is no simple closed form solution for the mean time to absorption, but we can arrive at the approximation as follows. We write $p = \frac{i}{N}$ for the proportion of a-alleles in the population and suppose that the mean time to absorption starting from p can be approximated by a twice differentiable function of p that we denote $t(p)$.

Now if the current number of alleles is i, the next generation has a number which is Binom$(N, \frac{i}{N})$ distributed. So the expected number of a alleles in the next generation is $N \cdot \frac{i}{N} = i$ and the variance is $N \cdot \frac{i}{N} \cdot \frac{N-i}{N} = \frac{i(N-i)}{N}$. the *change*, δp, in the *proportion* of a alleles then has mean zero and variance $\frac{1}{N^2}\frac{i(N-i)}{N} = \frac{1}{N}p(1-p)$. In particular, we expect δp to be *small*.

Using the argument that we used to justify the system (2),

$$\begin{aligned} t(p) &= \sum_{\delta p} \mathbb{P}[p \mapsto p + \delta p](t(p + \delta p) + 1) \\ &\approx \sum_{\delta p} \mathbb{P}[p \mapsto p + \delta p](t(p) + \delta p t'(p) + \frac{(\delta p)^2}{2} t''(p) + 1) \\ &= t(p) + t'(p)\mathbb{E}[\delta p] + \frac{1}{2}t''(p)\mathbb{E}[(\delta p)^2] + 1, \end{aligned}$$

and from what we just said

$$\mathbb{E}[\delta p] = 0, \quad \mathbb{E}[(\delta p)^2] = \frac{1}{N}p(1-p).$$

Substituting,

$$t(p) = t(p) + \frac{1}{2N}p(1-p)t''(p) + 1$$

or

$$p(1-p)t''(p) = -2N, \quad t(0) = t(1) = 0.$$

This can be solved to give (3).

Now let's turn to the Moran model.

Example 1.9 *In the Moran model, if the initital frequency of a alleles is p then the expected time to absorption is approximately*

$$\tau(p) = -2\left(p \log p + (1-p)\log(1-p)\right).$$

Recall that in our Moran model we are already working in 'coalescent time', corresponding to measuring time in units of size N in the Wright-Fisher model, so for large populations the absorption times in the two models are approximately the same.

To calculate the mean time to absorption in the Moran model we first calculate the mean number of transitions of the embedded discrete time chain until absorption and then multiply by the expected time between transitions. Again the absorbing states are $\{0, N\}$.

Consider then the embedded chain. Substituting into equation (2) from Theorem 1.7 we obtain (suppressing A in our notation)

$$\begin{aligned} k_i &= 1 + \sum_j p_{ij} k_j \\ &= 1 + \frac{i(N-i)}{N^2}k_{i+1} + \frac{i(N-i)}{N^2}k_{i-1} + \left(1 - \frac{2i(N-i)}{N^2}\right)k_i. \end{aligned}$$

Rearranging

$$k_{i+1} - 2k_i + k_{i-1} = -\frac{N^2}{i(N-i)} \quad i = 1, \ldots, N-1$$

and of course $k_0 = k_N = 0$. This can be solved (see the problem sheet) to yield

$$k_i = N \left\{ \sum_{j=1}^{i} \frac{N-i}{N-j} + \sum_{j=i+1}^{N-1} \frac{i}{j} \right\}.$$

This tells us the expected number of transitions of the embedded chain before absorption. The time between transitions is exponentially distributed with parameter $\binom{N}{2}$ and so has mean $\frac{2}{N(N-1)}$. Thus the expected time to absorption is

$$= \frac{2}{N-1} \left\{ \sum_{j=1}^{i} \frac{N-i}{N-j} + \sum_{j=i+1}^{N-1} \frac{i}{j} \right\}.$$

Now again we are really interested in large populations, corresponding to large N, and then writing $i = pN$ we see that the time to absorption is

$$\begin{aligned} \frac{2N}{N-1} \left\{ \sum_{j=1}^{pN} \frac{(1-p)}{N-j} + \sum_{j=pN+1}^{N-1} \frac{p}{j} \right\} &\approx 2 \left\{ (1-p) \sum_{j=1}^{pN} \frac{1}{N-j} + p \sum_{j=pN+1}^{N-1} \frac{1}{j} \right\} \\ &\approx -2 \left(p \log p + (1-p) \log(1-p) \right) \end{aligned}$$

as required. □

Recall that in our Moran model we are already working in 'coalescent time' and so this corresponds exactly to the approximation obtained for the Wright-Fisher model.

2 The infinite population limit

For large populations we have seen that the Wright-Fisher and Moran models have approximately the same genealogy (if we measure time in appropriate units in the Wright-Fisher model) - so we have the freedom to choose which model to use. However, even in this biologically very simple setting of no mutation or selection exact calculations for the Wright-Fisher model are impossible and the Moran model, though more tractable, still leads to rather complex expressions. The Moran model retains some of that tractability when we introduce more evolutionary forces such as mutation and selection, because it will still be a birth and death process and so many exact expressions are still available. However, these expressions are generally so complex as to completely obscure the *effects* of the different evolutionary parameters and so one would like a simpler model still - at least for large populations and our next aim is to obtain such a model by passing to an *infinite* population limit.

Since the basic objects of study in population genetics are the genealogical trees that relate individuals in a sample from the population, we should like the distribution of these genealogical trees to be preserved as we pass to the infinite population limit. The work that really explained why this can be done in a cast iron mathematical formalism is due to Peter Donnelly and Tom Kurtz. We're only going to see the vaguest of outlines of their powerful work here, but let's try to understand the basic idea.

Recall our graphical representation of the Moran model. Each pair of lines had an exponential rate one clock associated to it and when the clock rang an arrow was drawn to represent one individual dying and the other reproducing. We convinced ourselves that the genealogical trees in this model were precisely those given by Kingman's coalescent. Now the labelling of individuals in our population was arbitrary - we could have taken any permutation of these labels and arrived at a process with exactly

the same distribution. The idea is to exploit this to choose a convenient labelling. What is crucial is that we can do this in such a way that the Nth population process is embedded in the $(N+k)$th for all $k \geq 1$.

First consider $N = 2$. The convention of Donnelly and Kurtz is to draw time horizontally and so we do that here. In our original graphical representation of the Moran model we would have drawn arrows in either direction (pointing up or down) at exponential rate one. We're now going to insist that all arrows point down. The time to the most recent common ancestor will still be exponential with parameter one, but the type of that ancestor will necessarily be the type of the individual at the lower level. Provided that the types of individuals at time zero were allocated in such a way that the distribution was the same even if we permuted the labels, then the distribution at time t of types under this model is the same as under our original labelling in the Moran model.

Now add another individual at 'level 3'. Again in our original graphical representation there would have been arrows both up and down between each pair of levels, now we insist that all arrows point downwards. So arrows emanate from level 3 at a total rate of 2 (one for arrows to level one and one for arrows to level two). Notice that the time until the first merger of ancestral lines is the minimum of the three exponential random variables that we have and that it is equally likely to be any of the three pairs of lineages that is involved in the merger. Once again, provided that the original allocation of types is such that the labels don't matter, the distribution of types at any future time t under this 'lookdown' model is the same as under our original Moran model. All we have done is change the labels.

In general we have the following picture. Recall that the stochastic process that records the times at which an exponential clock with parameter λ rings - that is the intervals between times are independent $Exp(\lambda)$ random variables - is called a Poisson process with rate λ.

Definition 2.1 (The N-particle lookdown process) *The N-particle lookdown process will be denoted by the vector $(\zeta_1(t), \ldots, \zeta_N(t))$. Each index is thought of as representing a 'level'. The evolution of the process is described as follows. The individual at level k is equipped with an exponential clock with rate $(k-1)$, independent of all other individuals. At the times determined by the corresponding Poisson process she selects a level uniformly at random from $\{1, 2, \ldots, k-1\}$ and adopts the current type of the individual at that level. (Her level does not change.)*

Remark 2.2 *Since the minimum of $(k-1)$ independent $Exp(1)$ random variables is $Exp(k-1)$, the rate in the Poisson process is exactly that dictated by there being a lookdown event between any two levels at rate one. When such an event takes place it is equally likely to be any of the $(k-1)$ $Exp(1)$ clocks that has rung, hence the uniform selection from levels $1, \ldots, k-1$.*

Notice that the N-particle lookdown process is embedded in the $(N+k)$-particle lookdown process for each $k \geq 1$. Moreover, since we already said that the genealogy of a sample of size n from the Moran model is an n-coalescent and since we've seen that the genealogy of the first n levels in the lookdown process is also an n-coalescent, with this labelling we have a nice consistent way of sampling from a Moran model of arbitrary size.

The genealogy of the sample is that of the first n levels in the lookdown process. And the evolution of those levels does not depend on the population size - because we only ever 'look down' we don't see the population size N at all. Indeed it could be arbitrarily large. So can we make sense of taking the limit as $N \to \infty$? The substance of Donnelly and Kurtz's paper from 1996 is that we can. Our next task is to identify the limiting population model.

Evidently if the population size is infinite it is not going to make sense to talk about the 'number' of a-alleles, but we can talk about the proportion. We're going to find a way to characterise the way that proportion evolves in the infinite population limit. To do this we need one more concept.

We'll call a continuous time Markov process $\{X_t\}_{t\geq 0}$ *time homogeneous* if the probability of a given transition in the time interval $(t, t+s)$ is independent of t. (Transitions can still depend on X_t, but if $X_t = x$ say, then they depend on X_t only through x, not through the time t.)

Definition 2.3 (Generator of a continuous time Markov process) *Let $\{X_t\}_{t\geq 0}$ be a real-valued continuous time Markov process. For simplicity suppose that it is time homogeneous. For a function $f : \mathbb{R} \to \mathbb{R}$ define*

$$\mathcal{L}f(x) = \lim_{\delta t \downarrow 0} \frac{\mathbb{E}[f(X_{\delta t}) - f(x)|X_0 = x]}{\delta t}$$

if the limit exists. We'll call the set $\mathcal{D}(\mathcal{L})$ of functions for which the limit exists the domain *of $\mathcal{L}$ and the operator $\mathcal{L}$ acting on $\mathcal{D}(\mathcal{L})$ the* infinitesimal generator *of $\{X_t\}_{t\geq 0}$.*

The point is that if I know $\mathcal{L}$ then I can write down a differential equation for the way that $\mathbb{E}[f(X_t)]$ evolves with time. If $\mathcal{L}f$ is defined for sufficiently many different functions then this completely characterises the distribution of $\{X_t\}_{t\geq 0}$.

Example 2.4 (The Moran model) *Suppose that a population of N genes evolves according to the Moran model and write $\{p_t\}_{t\geq 0}$ for the stochastic process that records the proportion of a alleles. Then*

$$\mathcal{L}f(p) = \binom{N}{2} p(1-p)\left(f(p+\frac{1}{N}) - f(p)\right) + \binom{N}{2} p(1-p)\left(f(p-\frac{1}{N}) - f(p)\right).$$

The transitions of the Moran model take place at the points of a Poisson process with rate $\binom{N}{2}$ and at the time of such a transition, the change in the proportion of a alleles is determined by the embedded chain given in Example 1.6. Thus if the proportion of a alleles is initially p, then at the time of the first transition

$$\begin{aligned} p &\mapsto p + \frac{1}{N} \quad \text{with probability } p(1-p), \\ p &\mapsto p - \frac{1}{N} \quad \text{with probability } p(1-p) \end{aligned}$$

and there is no change with probability $1 - 2p(1-p)$. The chance that we see a transition in a time interval of length δt is

$$\binom{N}{2}\delta t + \mathcal{O}((\delta t)^2)$$

and the probability of seeing more than one transition is $\mathcal{O}((\delta t)^2)$. Putting all this together gives that for $f : [0,1] \to \mathbb{R}$ and $p = \frac{i}{N}$ for some $i \in \{0, 1, \ldots, N\}$

$$\begin{aligned} \mathcal{L}f(p) &= \lim_{\delta t \downarrow 0} \frac{1}{\delta t}\Big\{ \binom{N}{2}\delta t \left[p(1-p)f(p+\frac{1}{N}) + p(1-p)f(p-\frac{1}{N}) + (1-2p(1-p))f(p)\right] \\ &\qquad + \left(1 - \binom{N}{2}\delta t\right) f(p) + \mathcal{O}((\delta t)^2) - f(p)\Big\} \\ &= \binom{N}{2} p(1-p)\left(f(p+\frac{1}{N}) - f(p)\right) + \binom{N}{2} p(1-p)\left(f(p-\frac{1}{N}) - f(p)\right). \end{aligned}$$

To see what our population process will look like for large N we take f to be twice continuously differentiable and use Taylor's Theorem to find an approximation for $\mathcal{L}f$. Thus

$$\begin{aligned} \mathcal{L}f(p) &= \binom{N}{2} p(1-p)\Big(f(p) + \frac{1}{N}f'(p) + \frac{1}{2N^2}f''(p) + \mathcal{O}(\frac{1}{N^3}) - f(p) \\ &\qquad + f(p) - \frac{1}{N}f'(p) + \frac{1}{2N^2}f''(p) + \mathcal{O}(\frac{1}{N^3}) - f(p)\Big) \\ &= \frac{1}{2}p(1-p)f''(p) + \mathcal{O}(\frac{1}{N}). \end{aligned}$$

It is reasonable to guess then that for the infinite population limit,

$$\frac{d}{dt}\mathbb{E}[f(p_t)|p_0 = p]\Big|_{t=0} = \frac{1}{2}p(1-p)f''(p).$$

The next question is, is there a Markov process for which this is true? The answer, it turns out, is yes. The limiting process is a well-defined continuous time, continuous space Markov process. In the lookdown process, p_t corresponds to the limiting proportion of a-alleles at time t as we let the number of levels tend to infinity. The lookdown construction allows simultaneous construction of the limiting process $\{p_t\}_{t\geq 0}$ *and* the genealogical trees relating individuals in the population. Thus although it doesn't really make sense to talk about individuals in our model for proportions, the lookdown construction says that we can still think of them as being there and moreover their genealogy is governed by Kingman's coalescent.

The limiting process $\{p_t\}_{t\geq 0}$ is called the Fisher-Wright diffusion and now we're going to step back from genetics for a while to learn a little about diffusions and how to calculate quantities relating to them.

3 Diffusions

A diffusion process $\{X_t\}_{t\geq 0}$ is a continuous space and time Markov process which traces out a continuous path as time evolves. So at any instant in time it is a continuous random variable but also any realisation of $\{X_t\}_{t\geq 0}$ is a continuous function of time.

Definition 3.1 *The transition density function of $\{X_t\}_{t\geq 0}$ is the function $p : \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ for which*

$$\mathbb{P}[X_t \in A | X_0 = x] \equiv \mathbb{P}_x[X_t \in A] = \int_A p(t, x, y)dy$$

for any subset $A \subseteq \mathbb{R}$.

The existence of such a function is guaranteed for all the processes that we consider here, but the proof of that is beyond our scope. Notice that $p(t, x, y)$ is just the probability density function for the position of X_t given that $X_0 = x$. One can consider more general processes, but we consider only *time-homogeneous* diffusions so that

$$\mathbb{P}[X_{t+s} \in A | X_s = x] = \mathbb{P}[X_t \in A | X_0 = x] \quad \forall s \geq 0, \forall x \in \mathbb{R}.$$

The Markov property tells us that for $s < t$

$$\mathbb{P}_x[X_t \in A | \{X_r\}_{0\leq r\leq s}] = \mathbb{P}_x[X_t \in A | X_s].$$

A useful consequence of this is the following.

Lemma 3.2 (The Chapman-Kolmogorov equation) *For $s < t$, $x, z \in \mathbb{R} \times \mathbb{R}$*

$$p(t, x, z) = \int_{\mathbb{R}} p(t - s, x, y)p(s, y, z)dy.$$

"Justification": For any set $A \subseteq \mathbb{R}$, an extension of the Partition Theorem tells us that

$$\mathbb{P}[X_t \in A] = \int_{\mathbb{R}} \mathbb{P}[X_t \in A | X_s = y]\mathbb{P}[X_s \in [y, y + dy)]$$

and since A is arbitrary the result follows.

The most fundamental example of a diffusion process is the process known as Brownian motion. The simplest way to think of Brownian motion is as a rescaling limit of random walk. So let $\{Z_i\}_{i\in\mathbb{N}}$ be independent identically distributed random variables with $\mathbb{P}[Z_i = 1] = \frac{1}{2} = \mathbb{P}[Z_i = -1]$, and let $S_n = \sum_{i=1}^n Z_i$. Taking $S_0 = 0$, $\{S_n\}_{n\geq 0}$ is the process known as simple random walk on $\mathbb{Z}$.

Now consider the rescaled process (in continuous time)

$$B_t^{(n)} = \frac{1}{\sqrt{n}} S_{[nt]}$$

where $[nt]$ denotes the integer part of nt. The rescaled process $B_t^{(n)}$ is again a random walk, but it takes steps at time intervals of length $\frac{1}{n}$ and the size of each step is $\pm\frac{1}{\sqrt{n}}$.

As $n \to \infty$, by the Central Limit Theorem

$$B_t^{(n)} \to B_t \sim N(0, t).$$

Moreover, since the steps taken by the random walk in disjoint time intervals are independent, $B_t^{(n)} - B_s^{(n)}$ is independent of $B_s^{(n)} - B_0^{(n)}$ for $0 < s < t$ and this independence is inherited by the limiting process. And of course, again by the Central Limit Theorem, $B_t - B_s \sim N(0, t - s)$. This is enough to uniquely identify the process and so we make the following definition.

Definition 3.3 (Brownian motion) *The real-valued stochastic process $\{B_t\}_{t\geq 0}$ is a* Brownian motion *if*

1. *For each $t > 0$ and $s \geq 0$, $B_{t+s} - B_s$ has the normal distribution with mean zero and variance $\sigma^2 t$ for some constant σ.*
2. *For each $n \geq 1$ and any times $0 \leq t_1 \leq \cdots \leq t_n$, the random variables $\{B_{t_r} - B_{t_{r-1}}\}_{r=1}^n$ are independent.*
3. *$B_0 = 0$.*
4. *B_t is continuous in $t \geq 0$.*

Remark 3.4

1. *The third condition is a convention. To construct Brownian motion starting from a point x just take $x + B_t$.*
2. *Our derivation from simple random walks led us to* standard *Brownian motion in which $\sigma^2 = 1$. The Brownian motion with variance parameter σ^2 described above is just a time change. If $\{B_t\}_{t\geq 0}$ is a standard Brownian motion, then $\{B_{\sigma^2 t}\}_{t\geq 0}$ is a Brownian motion with variance parameter σ^2.*
3. *Continuity of the paths of Brownian motion is in some sense a consequence of the first three conditions, but we should like to specify once and for all that our Brownian motion has continuous paths.*
4. *Although the paths of Brownian motion are continuous, this does not mean that they are in any other sense nice. For example they are nowhere differentiable. For intuition it is probably best to think about infinitesimal random walks.*

One can also introduce a drift - or a bias - into the Brownian motion. If we replace the first condition by $B_{t+s} - B_s$ is normally distributed with mean μt and variance $\sigma^2 t$ we arrive at Brownian motion with drift. It is easy to check that this process is just $\mu t + B_t$ at time t.

In a general diffusion, the parameters μ and σ^2 are allowed to depend on the spatial positon of the process. We assume that they are sufficiently nicely behaved that over an infinitesimal neighbourhood of the point x they are approximately constant. Then for an infinitesimal time increment δt we expect the increment of the diffusion, which we denote by δX, to be approximately $N(\mu(x)\delta t, \sigma^2(x)\delta t)$ distributed so that locally the process behaves like Brownian motion plus drift. Then

$$\mathbb{E}[\delta X] = \mu(x)\delta t, \quad var(\delta X) = \sigma^2(x)\delta t$$

(and so $\mathbb{E}[(\delta X)^2] = \sigma^2(x)\delta t + \mathcal{O}((\delta t)^2)$ and $\mathbb{E}[(\delta X)^k] = \mathcal{O}((\delta t)^2)$ for all $k \geq 3$). We don't actually require normally distributed increments for what follows, these consequences suffice.

Example 3.5 *1. For standard Brownian motion $\mu = 0$, $\sigma^2 = 1$.*

2. For the diffusion limit of the Moran model (reverting to our notation p in place of x), $\mathbb{E}[p(t+\delta t) - p(t)] = 0$ (since the positive and negative increments in p are equally probable) and so $\mu \equiv 0$, whereas

$$\mathbb{E}[(p(t+\delta t) - p(t))^2] = \binom{N}{2}\delta t \cdot 2p(1-p)\frac{1}{N^2} \approx p(1-p)\delta t,$$

so $\sigma^2(p) = p(1-p)$.

So given the 'drift' and 'diffusion' (or 'variance') coefficients $\mu(x)$ and $\sigma^2(x)$, can we say anything about the transition densities for the diffusion? That is can we pass from knowledge of the local (infinitesimal) behaviour to knowledge of the global?

We are going to *assume* that the transition densities exist and are twice continuously differentiable in t. This is certainly true if $\mu(x)$ and $\sigma^2(x)$ are smooth functions of x, but the proof of this is beyond our (analytic) scope.

We're going to write (a, b) for the state space of the diffusion. For Brownian motion this is $\mathbb{R}$, but for many of our genetic examples it will be $(0, 1)$.

Our first calculation is analogous to that performed for the Moran model at the end of §2. Let us write

$$u(t, x) = \mathbb{E}[f(X_t)|X_0 = x]$$

where f is a fixed real-valued function.

Since the diffusion has continuous paths, $u(0+, x) = f(x)$. We're going to establish a differential equation for $u(t, x)$. Our starting point is the Chapman-Kolmogorov equation.

$$\begin{aligned}
u(t+\delta t, x) &= \mathbb{E}[f(X_{t+\delta t})|X_0 = x] \\
&= \int p(t+\delta t, x, z) f(z) dz \\
&= \int \left(\int p(\delta t, x, y) p(t, y, z) dy \right) f(z) dz \\
&= \int p(\delta t, x, y) \left(\int p(t, y, z) f(z) dz \right) dy \\
&= \int p(\delta t, x, y) u(t, y) dy.
\end{aligned}$$

Now let's rewrite y rather more suggestively as $x + \delta x$. Then the right hand side above becomes

$$\begin{aligned} u(t+\delta t, x) &= \int p(\delta t, x, x+\delta x) u(t, x+\delta x) d(\delta x) \\ &= \int p(\delta t, x, x+\delta x) \left\{ u(t,x) + \delta x u'(t,x) + \frac{(\delta x)^2}{2} u''(t,x) + \mathcal{O}((\delta x)^3) \right\} d(\delta x) \\ &= u(t,x) + u'(t,x)\mathbb{E}_x[\delta X] + \frac{1}{2} u''(t,x)\mathbb{E}_x[(\delta X)^2] + \mathcal{O}(\mathbb{E}[(\delta X)^3]). \end{aligned}$$

So

$$\frac{u(t+\delta t, x) - u(t,x)}{\delta t} = \mu(x) u'(t,x) + \frac{1}{2}\sigma^2(x) u''(t,x) + \mathcal{O}(\delta t).$$

Letting $\delta t \to 0$ yields *Kolmogorov's backward equation*,

$$\frac{\partial u}{\partial t} = \mu u' + \frac{1}{2}\sigma^2 u''. \tag{4}$$

Of course, setting $t = 0$, $u(0, x) = f(x)$.

Since this equation holds for all choices of f we can deduce that it is also true for $p(t, x, y)$, so we have

Lemma 3.6

$$\frac{\partial}{\partial t} p(t,x,y) = \mu(x) \frac{\partial}{\partial x} p(t,x,y) + \frac{1}{2}\sigma^2(x) \frac{\partial^2}{\partial x^2} p(t,x,y) \tag{5}$$

and $p(0, x, y) = \delta_x$.

The object δ_x is not a *function*, it is what I'd call a point mass - often called the Dirac delta function. It's just defined by

$$\int f(y)\delta_x(y)dy = f(x)$$

for all functions f. The reason that equation (5) is referred to as a *backward* equation is because it tells us about the relationship between t and the *initial* point for the Markov process, x. From the point of view of X_t this is backwards.

In some settings it is natural to think about the relationship between t and y. To uncover such a relationship we once again use the Chapman-Kolmogorov equation. Note that

$$p(t+s, x, z) = \int p(s, x, y) p(t, y, z) dy \tag{6}$$

depends on s and t only through $s + t$ so the derivative with respect to s is the same as the derivative with respect to t. Applying this to the right hand side of (6) gives

$$\begin{aligned} \int \frac{\partial p}{\partial s}(s,x,y) p(t,y,z) dy &= \int p(s,x,y) \frac{\partial p}{\partial t}(t,y,z) dy \\ &= \int p(s,x,y) \big\{ \mu(y) \frac{\partial p}{\partial y}(t,y,z) + \frac{1}{2}\sigma^2 \frac{\partial^2 p}{\partial y^2}(t,y,z) \big\} dy. \end{aligned}$$

We now integrate by parts and use $'$ to denote differentiation with respect to y to obtain

$$\begin{aligned}\int \frac{\partial p}{\partial s}(s,x,y)p(t,y,z)dy &= [p(s,x,y)\mu(y)p(t,y,z)]_a^b - \int (\mu(y)p(s,x,y))' p(t,y,z)dy \\ &\quad + [\frac{1}{2}\sigma(y)^2 p(s,x,y)\frac{\partial p}{\partial y}(t,y,z)]_a^b - \int \frac{1}{2}\left(\sigma(y)^2 p(s,x,y)\right)' \frac{\partial p}{\partial y}(t,y,z)dy \\ &= \Big[p(s,x,y)\mu(y)p(t,y,z) + \frac{1}{2}\sigma(y)^2 p(s,x,y)\frac{\partial p}{\partial y}(t,y,z) \\ &\qquad - \frac{1}{2}\left(\sigma(y)^2 p(s,x,y)\right)' p(t,y,z)\Big]_a^b \\ &\quad - \int (\mu(y)p(s,x,y))' p(t,y,z)dy + \int \frac{1}{2}\left(\sigma(y)^2 p(s,x,y)\right)'' p(t,y,z)dy.\end{aligned}$$

Now if the boundary terms vanish, this gives

$$\int \frac{\partial p}{\partial s}(s,x,y)p(t,y,z)dy = \int \{ -(\mu(y)p(s,x,y))' p(t,y,z) + \frac{1}{2}\left(\sigma(y)^2 p(s,x,y)\right)'' \}p(t,y,z)dy.$$

Taking $t = 0$ (or rather letting $t \downarrow 0$ we deduce that

Lemma 3.7

$$\frac{\partial p}{\partial s}(s,x,y) = -\frac{\partial}{\partial y}\left(\mu(y)p(s,x,y)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma(y)^2 p(s,x,y)\right). \tag{7}$$

Equation (7) is known as the Kolmogorov *forward* equation.

What about the boundary terms that we have neglected? In most cases of practical interest - including all the ones that we'll encounter in this course - they can be ignored. To see why, suppose that z is not a boundary point and consider what happens to $p(t,y,z)$ as $t \to 0$. We already said that this will be a delta-function at y in the limit. So if y is a boundary point, $p(t,y,z) \to 0$ for z not in the boundary. Similarly, its derivative will vanish (provided that our coefficients $\mu(x)$ and $\sigma(x)^2$ are not *too* horrible.

Again *forward* equation refers to the fact that we have established a relationship between t and the density function for the Markov process at time t at the point y - so forwards in time for the Markov process.

Notice that the forwards equation is meaningless unless $\mu(x)$ and $\sigma^2(x)$ are differentiable. There is no such restriction for our backwards equation.

Before exploring the ramifications of the backwards equation, we're now going to look at some things that we can calculate from the forwards equation without actually explicitly calculating the transition probabilities.

Example 3.8 *Consider a simple model for population growth: the simple birth and death process. Each individual in a large population is equipped with two exponential clocks, one of rate λ, one of rate μ. If the rate μ clock rings first, then the individual dies, whereas if the rate λ clock rings first then the individual splits into two. The offspring go on to evolve independently of one another in the same way as their parent. Assuming that the population is large, how does its size evolve with time.*

Solution. We are going to suppose that the population size is measured in terms of very large units N and look at its evolution over very long timescales which we'll also measure in units of size N. Unless $\lambda - \mu$ is $\mathcal{O}(\frac{1}{N})$ (in these units) the population will either die out very quickly or grow very quickly, so assume that $N(\lambda - \mu) = b$ and $\lambda + \mu = 2a + \mathcal{O}(\frac{1}{N})$ for some constants a and b. Let's consider what a diffusion approximation to this model would look like.

So denote the population size at time t by X_t. Then there are NX_t individuals alive at time t. Since we are measuring time in units of size N, the chance for each individual that their 'death clock' goes off in the next δt of time is $N\mu\delta t$ and similarly the probability that their 'birth clock' goes off is $N\lambda\delta t$. So if we write ΔX for the change in population size over the next δt of time, the chance that $\Delta X = -\frac{1}{N}$ is $NX_t \cdot N\mu\delta t + \mathcal{O}(\delta t^2)$ and the chance that $\Delta X = \frac{1}{N}$ is $NX_t \cdot N\lambda\delta t + \mathcal{O}(\delta t^2)$. The chance of a change bigger than this is $\mathcal{O}(\delta t^2)$. Thus

$$\begin{aligned}\mathbb{E}[\Delta X] &= \frac{1}{N}NX_t \cdot N\lambda\delta t - \frac{1}{N}NX_t \cdot N\mu\delta t + \mathcal{O}(\delta t^2)\\ &= N(\lambda-\mu)X_t\delta t + \mathcal{O}(\delta t^2)\\ &= bX_t\delta t + \mathcal{O}(\delta t^2).\end{aligned}$$

$$\begin{aligned}\mathbb{E}[(\Delta X)^2] &= \frac{1}{N^2}NX_t \cdot N\lambda\delta t + \frac{1}{N^2}NX_t \cdot N\mu\delta t + \mathcal{O}(\delta t^2)\\ &= 2aX_t\delta t + \mathcal{O}(\delta t^2).\end{aligned}$$

The diffusion approximation will therefore have infinitesimal mean and variance bX_t and $2aX_t$ resepctively and the corresponding backward equation is

$$\frac{\partial}{\partial t}p(t,x,y) = bx\frac{\partial}{\partial x}p(t,x,y) + ax\frac{\partial^2}{\partial x^2}p(t,x,y).$$

The *forwards* equation is

$$\frac{\partial}{\partial t}p(t,x,y) = -\frac{\partial}{\partial y}\left(yp(t,x,y)\right) + \frac{\partial^2}{\partial y^2}\left(ayp(t,x,y)\right). \tag{8}$$

Let's write $M(t)$ for the expected population size at time t, that is

$$M(t) = \int_0^\infty yp(t,x,y)dy.$$

Multiplying both sides of (8) by y and integrating over $[0,\infty)$ we have

$$M'(t) = -b\int_0^\infty y\frac{\partial}{\partial y}\left(yp(t,x,y)\right)dy + a\int_0^\infty y\frac{\partial^2}{\partial y^2}\left(yp(t,x,y)\right)dy.$$

Now integrate by parts on the right hand side.

$$\begin{aligned}M'(t) &= -b\left[y^2p(t,x,y)\right]_0^\infty + b\int_0^\infty yp(t,x,y)dy\\ &\qquad + a\left[y\frac{\partial}{\partial y}\left(yp(t,x,y)\right)\right]_0^\infty - a\int_0^\infty \frac{\partial}{\partial y}\left(yp(t,x,y)\right)dy\\ &= bM(t) - a\left[yp(t,x,y)\right]_0^\infty\\ &= bM(t).\end{aligned}$$

We have used the fact that $p(t,x,y) \to 0$ very fast at infinity. This can be proved, but for this course we take it on trust. The point is that since our 'drift' and 'variance' are both finite, the chance of getting to y in time t decays exponentially as $t \to \infty$. (You can cite this result in the course if you need it.)

So $M'(t) = bM(t)$ or

$$M(t) = M(0)e^{bt}.$$

Similar manipulations give that the variance in the population size is proportional to $\frac{2a}{b}e^{bt}(e^{bt} - 1)$. Notice that at no point did we actually calculate $p(t, x, y)$ itself. □

Before exploring some of the consequences of the backward equation, let's do one more example with the forward equation. First we need to recall another notion from Markov chains. For a Markov chain X with transition matrix P, let us write

$$\begin{aligned}\pi_j^{(n+1)} &= \mathbb{P}[X_{n+1} = j] \\ &= \sum_i \mathbb{P}[X_{n+1} = j | X_n = i]\mathbb{P}[X_n = i] \\ &= \sum_i p_{ij}\pi_i^{(n)},\end{aligned}$$

so that $\pi^{(n+1)} = \pi^{(n)}P$.

In many examples as $n \to \infty$ the number of visits to each site before n, as a proportion of n, of a typical realisation of the chain converges to a deterministic limit obtained by solving

$$\pi = \pi P$$

and normalising so that the sum of the entries in π is 1. The resulting vector of probabilities, π, is called a *steady state* or *stationary distribution* for the system. Results which relate steady state probabilities to frequency of visits in realisations of Markov chains are called *ergodic theorems*. If there is a unique steady state then, no matter what the initial condition, the chain eventually settles down to that steady state in the sense that if I look at a very large time, the probabilities for being in the different states are given by the vector π.

Ergodic theorems are certainly not valid for all Markov chains. There are essentially two things that can go wrong:

1. The first is obvious - they can have 'traps'. For example, the states can split into distinct groups in such a way that the system cannot get from one group to another. So we require that the chain has the property that every state can be reached from every other state (in one or more transitions). Such a chain is said to be *irreducible*.

2. The second barrier is more subtle. It may for example be the case that for some initial states, the possible states split up into two groups - one visited only at even times and the other only at odd times. (This is the case for simple random walk on $\mathbb{Z}$.) In general there may be states that are only visited by the chain at times divisible by some integer k. A Markov chain with this property is said to be *periodic*. We require that the chain is *aperiodic* - that is there are no states with this property for any $k = 2, 3, \ldots$.

Notice that our basic Moran and Wright-Fisher models both get trapped in 0 and N, so that they are not irreducible. However, if we add mutation between types then there is a stationary distribution and in the case of the Moran model it can even be found explicitly.

Theorem 3.9 *If a Markov chain with a finite number of states is irreducible and aperiodic then it has a unique steady state probability vector π such that $\pi = \pi P$. As $n \to \infty$ the probability vector $\pi^{(n)}$ tends to π independent of the initial vector $\pi^{(0)}$.*

We'd like to formulate something similar to this result for diffusion processes. Now we have a continuous random variable at each time, so it makes sense to consider not the probability mass function ($\pi^{(n)}$ in our Markov chain world) but rather the distribution function. So let's write

$$F(t,x,y) = \mathbb{P}[X_t \leq y | X_0 = x] = \int_{-\infty}^{y} p(t,x,u)du.$$

This plays the rôle of $\pi^{(n)}$ in the diffusion world and if there is an analogue of Theorem 3.9 then for very large times, irrespective of x, the distribution function of X_t should converge to $F(y) = \lim_{t\to\infty} F(t,x,y)$. Our aim now is to identify $F(y)$.

In the Markov chain world, π was fixed under the forwards in time evolution of the probabilities. In the diffusion world it is the Kolmogorov forward equation that tells us how $F(t,x,y)$ evolves with time. So recall the Kolmogorov's forwards equation

$$\frac{\partial}{\partial t}p(t,x,y) = -\frac{\partial}{\partial y}\left(\mu(y)p(t,x,y)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2(y)p(t,x,y)\right).$$

So now we integrate $\frac{\partial}{\partial t}p(t,x,u)$ with respect to u to obtain

$$\begin{aligned}
\frac{\partial}{\partial t}F(t,x,y) &= \int_{-\infty}^{y}\left\{-\frac{\partial}{\partial u}\left(\mu(u)p(t,x,u)\right) + \frac{1}{2}\frac{\partial^2}{\partial u^2}\left(\sigma^2(u)p(t,x,u)\right)\right\}du \qquad (9)\\
&= \left[-\mu(u)p(t,x,u)\right]_{-\infty}^{y} + \left[\frac{1}{2}\frac{\partial}{\partial u}\left(\sigma^2(u)p(t,x,u)\right)\right]_{-\infty}^{y}\\
&= -\mu(y)p(t,x,y) + \frac{1}{2}\frac{\partial}{\partial y}\left(\sigma^2(y)p(t,x,y)\right) + \text{ boundary terms.}
\end{aligned}$$

If the system settles down to a steady state, then $\frac{\partial}{\partial t}F(t,x,y) \to 0$. If we write $p(y)$ for the limiting density function (if it exists), so $p(y) = \lim_{t\to\infty} p(t,x,y)$, then

$$-\mu(y)p(y) + \frac{1}{2}\frac{d}{dy}\left(\sigma^2(y)p(y)\right) = C,$$

where C is a constant. It turns out that unless $C = 0$ we can't arrange that $\int p(y)dy = 1$, so take $C = 0$ in what follows. Then the equation is easily solved:

$$p(y) = \frac{\text{Const}}{\sigma^2(y)}\exp\left(2\int_{\eta}^{y}\frac{\mu(z)}{\sigma^2(z)}dz\right). \qquad (10)$$

We just fix any point η in the domain of X_t for the base of the indefinite integral and then fix the constant so that $\int p(y)dy = 1$. So if a stationary distribution exists for the diffusion, its density function will be given by (10). We did *not* need to find $p(t,x,y)$ at a finite time to determine this.

One way to think about the calculation that we just did, which was very popular in the 'classical' mathematical population genetics literature, is in terms of probability flux. The left hand side of (9) is the rate at which probability 'flows' from left to right through the point y, that is the 'probability flux' through y. In the stationary state the 'flux' is zero.

4 Speed and Scale

We're now going to turn to the ramifications of the backward equation. But in exploiting the backward equation we are going to turn to our advantage a technical point that so far we have washed over. Consider the backward equation

$$\frac{\partial u}{\partial t} = \mu(x)\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2} \qquad (11)$$

in an open interval (a, b) which may be finite or infinite. We have already seen that if $p(t, x, y)$ are the transition probability densities for our diffusion (with (a, b) a subset of - or equal to - the state space of the diffusion) then

$$u(t, x) = \int p(t, x, y)u(0, y)dy = \mathbb{E}[u(0, X_t)] \tag{12}$$

yields a solution to (11). If μ and σ^2 are sufficiently regular that the forwards equation makes sense, then there is a unique minimal function $p(t, x, \cdot)$ such that (12) yields a solution to (11). The catch is that for fixed t, x, the kernel $p(t, x, \cdot)$ may represent a *defective* distribution, that is it may not integrate to one. (The backward equation also has minimal solution $p(t, \cdot, y)$.) The backward equation only determines the process uniquely when the minimal solution is not defective. In all other cases, the nature of the process is determined by additional boundary conditions. These are most easily thought of in the context of simple random walk on $\mathbb{Z}_+$. If you think of 'gambler's ruin' in which a player repeatedly plays a game in which he wins £1 with probability p and loses £1 with probability $1 - p$, independently on each play, with the rule that he must stop playing when his fortune reaches £ 0, then 0 is an *absorbing barrier* for the Markov process which tracks his fortune. If instead the random walk is *instantaneously returned* to position 1 when it hits zero and the process continues forever, then 0 is a *reflecting barrier*.

Boundary conditions appear if and only if a boundary point can be reached - a well defined concept in diffusion processes because of continuity of sample paths. In some diffusion processes, with probability one, no boundary point is ever reached (this is the case for Brownian motion on $\mathbb{R}$). Then the minimal solution stands for a proper probability distribution and no other solutions exist. In all other cases, the minimal solution regulates the process until a boundary is reached. It corresponds to absorbing barriers, that is it describes a process that stops when a boundary point is reached. This is the most important type of process, not only because all other processes are extensions of it, but even more because all *first passage probabilities* can be calculated by artificially imposing absorbing barriers.

Here is the sort of question that we might want to solve in genetics: Suppose that our population has two alleles, a and A, with the initial frequency of a alleles given by p. What is the probability that the frequency of a alleles hits zero before it hits one? In other words, assuming that there are no new a alleles being produced by mutation, what is the probability that the a allele is eventually lost from the population?

We assume that the frequency of a alleles follows a diffusion process with absorbing barriers at zero and one. Write $P_0(p)$ for the probability of absorption at $X = 0$ if initially $X_0 = p$ and $P_1(p)$ for the corresponding probability of absorption at $X = 1$.

Let's write

$$F(t, p, x) = \int_0^x p(t, x, y)dy.$$

Integrating the backwards equation gives

$$\frac{\partial}{\partial t}\int_0^x p(t, p, y)dy = \mu(p)\int_0^x \frac{\partial}{\partial p}p(t, p, y)dy + \frac{1}{2}\sigma^2(p)\int_0^x \frac{\partial^2}{\partial p^2}p(t, p, y)dy,$$

that is

$$\frac{\partial}{\partial t}F(t, p, x) = \mu(p)\frac{\partial}{\partial p}F(t, p, x) + \frac{1}{2}\sigma^2(p)\frac{\partial^2}{\partial p^2}F(t, p, x).$$

Now we let $x = 0+$.

$$F(t, p, 0+) = \mathbb{P}[\text{at time } t, X \text{ has been absorbed at } 0] \equiv P_0(t, p).$$

In this notation,

$$\frac{\partial}{\partial t}P_0(t,p) = \mu(p)\frac{\partial}{\partial p}P_0(t,p) + \frac{1}{2}\sigma^2(p)\frac{\partial^2}{\partial p^2}P_0(t,p).$$

Now letting $t \to \infty$, $P_0(t,p) \to P_0(p)$ and $\frac{\partial}{\partial t}P_0(t,p) \to 0$ and, at least formally,

$$0 = \mu(p)\frac{dP_0}{dp}(p) + \frac{1}{2}\sigma^2(p)\frac{d^2P_0}{dp^2}(p),$$

with boundary conditions $P_0(0) = 1$ and $P_0(1) = 0$. The solution is easily obtained: we have a first order equation for $\frac{dP_0}{dp}(p)$ with solution

$$\frac{dP_0}{dp}(p) = \text{Const}\exp\left(-2\int^p \frac{\mu(z)}{\sigma^2(z)}dz\right)$$

and using the boundary conditions, provided that the denominator is finite we have

$$P_0(p) = \frac{\int_p^1 \exp\left(-2\int^y \frac{\mu(z)}{\sigma^2(z)}dz\right)dy}{\int_0^1 \exp\left(-2\int^y \frac{\mu(z)}{\sigma^2(z)}dz\right)dy}.$$

Similarly for $P_1(p)$ we obtain

$$P_1(p) = \frac{\int_0^p \exp\left(-2\int^y \frac{\mu(z)}{\sigma^2(z)}dz\right)dy}{\int_0^1 \exp\left(-2\int^y \frac{\mu(z)}{\sigma^2(z)}dz\right)dy}.$$

Notice in particular that if $\int_0^1 \exp\left(-2\int^y \frac{\mu(z)}{\sigma^2(z)}dz\right)dy$ is finite then $P_0(p) + P_1(p) = 1$. That is, the probability of going on for ever, never reaching 0 or 1 is zero.

Definition 4.1 *For a diffusion X_t on (a,b) with drift and variance μ and σ^2, the* scale function *is defined by*

$$S(x) = \int_{x_0}^x \exp\left(-\int_\eta^y \frac{2\mu(z)}{\sigma^2(z)}dz\right)dy,$$

where x_0, η are points fixed (arbitrarily) in (a,b).

Lemma 4.2 *If $a < a_0 < X_0 < b_0 < b$ then the probability of X_t hitting a_0 before b_0 is*

$$\frac{S(b_0) - S(X_0)}{S(b_0) - S(a_0)}.$$

Proof. Mimic what we did above but with a_0, b_0 'artificial' absorbing boundaries. □

Remark 4.3 *Notice that η cancels in the ratio and x_0 in the difference.*

This tells us the probability that we exit (a,b) for the first time through a, but can we glean some information about how long we must wait for X_t to exit the interval (a,b) (either through a or b) or, more generally, writing T^* for the first exit time of (a,b), can we say anything about $\mathbb{E}[\int_0^{T^*} g(X_s)ds|X_0 = p]$? (Putting $g = 1$ this gives the mean exit time.) Let us write

$$w(p) = \mathbb{E}[\int_0^{T^*} g(X_S)ds|X_0 = p]$$

and we'll derive the differential equation satisfied by w.

We assume that g is continuous. First note that $w(a) = w(b) = 0$. Now consider a small interval of time of length h. We're going to split the integral into the contribution up to time h and after time h. Because X_t has no memory, (more mathematically, because $\{X_t\}_{t\geq 0}$ has the Markov property),

$$\mathbb{E}[\int_h^{T^*} g(X_s)ds|X_h = z] = \mathbb{E}[\int_0^{T^*} g(X_s)ds|X_0 = z] = w(z)$$

and so for $a < p < b$

$$w(p) = \mathbb{E}[\int_0^h g(X_s)ds|X_0 = p] + \mathbb{E}[w(X_h)|X_0 = p]. \tag{13}$$

Since g is continuous and the paths of X are continuous we have the approximation

$$\mathbb{E}[\int_0^h g(X_s)ds|X_0 = p] = hg(p) + \mathcal{O}(h^2) \tag{14}$$

and just as in our derivation of the backward equation

$$\begin{aligned}
\mathbb{E}[w(X_h)|X_0 = p] &= \mathbb{E}[w(p + \Delta X)|X_0 = p] \\
&= \mathbb{E}[w(p) + \Delta X w'(p) + \frac{1}{2}(\Delta X)^2 w''(p) + \mathcal{O}(\Delta X^3)] \\
&= w(p) + \mu(p)hw'(p) + \frac{1}{2}\sigma^2(p)hw''(p) + \mathcal{O}(h^2).
\end{aligned} \tag{15}$$

Combining (15) and (14) with (13) we see that

$$\mu(p)w'(p) + \frac{1}{2}\sigma^2(p)w''(p) + g(p) = \mathcal{O}(h),$$

so that letting $h \to 0$, w satisfies

$$\mu(p)w'(p) + \frac{1}{2}\sigma^2(p)w''(p) = -g(p), \quad w(a) = 0 = w(b). \tag{16}$$

Let us now turn to solving this equation.

As in our derivation of the scale function it is convenient to fix $\eta \in (a, b)$. Then using an integrating factor, we rewrite (16) as

$$\frac{d}{dp}\left(\exp\left(\int_\eta^p \frac{2\mu(z)}{\sigma^2(z)}dz\right)w'(p)\right) = -\frac{2g(p)}{\sigma^2(p)}\exp\left(\int_\eta^p \frac{2\mu(z)}{\sigma^2(z)}dz\right).$$

Now recall that

$$S(x) = \int_a^x \exp\left(-\int_\eta^y \frac{2\mu(z)}{\sigma^2(z)}dz\right)dy$$

and let us write

$$m(x) = \frac{1}{\sigma^2(x)S'(x)},$$

that is

$$m(x) = \frac{1}{\sigma^2(x)}\exp\left(\int_\eta^x \frac{2\mu(z)}{\sigma^2(z)}dz\right),$$

then

$$\frac{d}{dp}\left(\frac{1}{S'(p)}w'(p)\right) = -2g(p)m(p)$$

and so

$$\frac{1}{S'(p)}w'(p) = -2\int_a^p g(\xi)m(\xi)d\xi + \beta$$

where β is a constant. Multiplying by $S'(p)$ and integrating gives

$$w(p) = -2\int_a^p S'(\xi)\int_a^\xi g(\eta)m(\eta)d\eta d\xi + \beta(S(p) - S(a)) + \alpha$$

for constants α, β. Since $w(a) = 0$, we immediately have that $\alpha = 0$. Reversing the order of integration,

$$\begin{aligned} w(p) &= -2\int_a^p\int_\eta^p S'(\xi)d\xi g(\eta)m(\eta)d\eta + \beta(S(p) - S(a)) \\ &= -2\int_a^p (S(p) - S(\eta))g(\eta)m(\eta)d\eta + \beta(S(p) - S(a)) \end{aligned}$$

and $w(b) = 0$ now gives

$$\beta = \frac{2}{S(b) - S(a)}\int_a^b (S(b) - S(\eta))g(\eta)m(\eta)d\eta.$$

Finally then

$$\begin{aligned} w(p) &= \frac{2}{S(b) - S(a)}\Big\{(S(p) - S(a))\int_a^b (S(b) - S(\eta))g(\eta)m(\eta)d\eta \\ &\qquad - (S(b) - S(a))\int_a^p (S(p) - S(\eta))g(\eta)m(\eta)d\eta\Big\} \\ &= \frac{2}{S(b) - S(a)}\Big\{(S(p) - S(a))\int_a^b (S(b) - S(\eta))g(\eta)m(\eta)d\eta \\ &\qquad + (S(b) - S(p))\int_a^p (S(\eta) - S(a))g(\eta)m(\eta)d\eta\Big\} \end{aligned}$$

where the last line is obtained by splitting the first integral into $\int_a^b = \int_p^b + \int_a^p$.

Now recall that $P_a(p)$, the probability of exit through a, is given by $\frac{S(b)-S(p)}{S(b)-S(a)}$ and $P_b(p)$, the probability of exit through b, is $\frac{S(p)-S(a)}{S(b)-S(a)}$ and so

$$w(p) = 2P_b(p)\int_p^b (S(b) - S(\eta))g(\eta)m(\eta)d\eta + 2P_a(p)\int_a^p (S(\eta) - S(a))g(\eta)m(\eta)d\eta.$$

Notice that this expression was found without ever explicitly calculating the transition densities for X_t. We have proved the following:

Theorem 4.4 *For a continuous function g,*

$$\mathbb{E}[\int_0^{T*} g(X_s)ds|X_0 = p] = \int_a^b G(p,\xi)g(\xi)d\xi,$$

where for $a < p < b$ we have

$$G(p,\xi) = \begin{cases} 2\frac{(S(p)-S(a))}{(S(b)-S(a))}(S(b) - S(\xi))m(\xi), & \text{for } p < \xi < b \\ 2\frac{(S(b)-S(p))}{(S(b)-S(a))}(S(\xi) - S(a))m(\xi), & \text{for } a < \xi < p, \end{cases}$$

with S the scale function given in Definition 4.1 and $m(\xi) = \frac{1}{\sigma^2(\xi)S'(\xi)}$.

Definition 4.5 *The function $G(p,\xi)$ is called the* Green's function *of the process X_t.*

By taking g to approximate $\mathbf{1}_{x_1,x_2}$ we see that $\int_{x_1}^{x_2} G(p,\xi)d\xi$ is the mean time spent by the process in (x_1,x_2) before exiting (a,b) if initially $X_0 = p$. Sometimes, the Green's function is called the *sojourn density.*

If a process has linear scale function, $S(\xi) = \xi + \text{Const}$, then its exit probabilities are exactly as for driftless Brownian motion. The function $m(\xi) = \frac{1}{\sigma^2(\xi)S'(\xi)}$ reduces to $\frac{1}{\sigma^2(\xi)}$ and can be thought of as a measure of the 'speed' of the process - by a random time change such a process becomes a Brownian motion and the time change is determined precisely by this $\frac{1}{\sigma^2(\xi)}$ which tells us how fast the clock should run. To see this heuristically, notice first that since

$$\mathcal{L}f(x) = \frac{d}{dt}\mathbb{E}[f(X_t)|X_0 = x]$$

if we perform the timechange $t \mapsto \alpha t$ for some constant α then by the chain rule, $\mathcal{L} \mapsto \alpha\mathcal{L}$. So if a process is in natural scale, since its generator is of the form $\frac{1}{2}\sigma^2(x)\frac{d^2}{dx^2}$, the timechange that will transform this into Brownian motion (whose generator is $\frac{1}{2}\frac{d^2}{dx^2}$) should locally look like $\frac{1}{\sigma^2(x)}$.

Definition 4.6 *The function $m(\xi) = \frac{1}{\sigma^2(\xi)}$ is the* density of the speed measure *or just the* speed density *of the process X_t.*

(Some textbooks define the speed measure to be twice this. We have followed Karlin & Taylor.) A fundamental fact about one-dimensional diffusions is that by first transforming space using the scale function, so considering $S(X_t)$ and then changing time via the speed density, we obtain a Brownian motion. Conversely, we can obtain a copy of the diffusion from Brownian motion by first changing time and then applying the inverse of the scale function.

Now that we are able to calculate quantities for the diffusion, let's check that in our genetic context it makes good predictions about our population models - at least for large populations.

Example 4.7 *Consider the Wright-Fisher diffusion with generator*

$$\mathcal{L}f(p) = \frac{1}{2}p(1-p)f''(p).$$

Notice that since it has no drift term ($\mu = 0$) it is already in natural scale, $S(x) = x + Const.$. What about $\mathbb{E}[T^*]$?

Using Theorem 4.4 with $g = 1$ we have

$$\begin{aligned}
\mathbb{E}_p[T^*] &= \mathbb{E}[\int_0^{T*} 1ds|X_0 = p] = \int_0^1 G(p,\xi)d\xi \\
&= 2\int_p^1 p(1-\xi)\frac{1}{\xi(1-\xi 0}d\xi + 2\int_0^p (1-p)\xi\frac{1}{\xi(1-\xi)}d\xi \\
&= 2p\int_p^1 \frac{1}{\xi}d\xi + 2(1-p)\int_0^p \frac{1}{1-\xi}d\xi \\
&= -2\left\{p\log p + (1-p)\log(1-p)\right\}
\end{aligned}$$

exactly as in Example 1.8. □

5 Selection and Mutation

The genetic models that we have considered so far have been very simple - all individuals in the population are equally fit and each individual passes on its exact genetic type to its offspring. Of course reality is not quite so straightforward. In this section we're going to increase the biological sophistication. The rôle of our Wright-Fisher and Moran models will now be to identify the appropriate diffusion approximation for large populations. Once we have obtained the diffusion approximation we can use the technology developed in the last section to make statements about the behaviour of the diffusion and hence about the approximate behaviour of large finite populations. Once we introduce these more realistic biological features the calculations become impossible in the Wright-Fisher model and although we can sometimes obtain expressions from the Moran model, these tend to be so complex as to obscure the real effects of the different genetic processes.

Any biologist would start from a Wright-Fisher model, so let's do the same thing here. For simplicity we're going to model N *haploid* individuals, that is each individual in our population carries exactly one copy of each chromosome so that an allele can be identified with a single unique parent in the previous generation. The extension to diploid populations like our own in which chromosomes are carried in pairs is discussed on the problem sheet. The extra ingredient is that the reproductive success of an allele - which is how we measure its fitness - can depend on which allelic type it is paired with in the diploid individual.

We are going to assume that the a alleles and A alleles have *relative fitnesses* $1+s:1$. What do we mean by this?

The basis of the Wright-Fisher model is that during reproduction each individual produces an essentially infinite number of offspring and it is from these offspring that the new generation is sampled. The proportion of types in the effectively infinite pool of offspring in the neutral world is exactly the same as in the parental generation, but in the selective world, if there are i a alleles and $N-i$ A alleles in the parental generation then the pool of potential offspring will have a proportion

$$\frac{(1+s)i}{(1+s)i+N-i}$$

of a alleles.

This accounts for selection, but we should also like to take into account mutation between types. So suppose that during the reproductive step each type a individual from the pool mutates to a type A with probability u and each type A mutates to a type a with probability v. then the proportion of the pool of potential offspring which is type a after selection and mutation is

$$\psi_i = \frac{(1+s)i}{(1+s)i+N-i}(1-u) + \frac{N-i}{(1+s)i+N-i}v.$$

From this pool of potential offspring we sample the next generation. So if the current number of a alleles is i (in our population of size N), then the next generation will contain exactly j a alleles with probability

$$p_{ij} = \binom{N}{j}\psi_i^j(1-\psi_i)^{N-j}.$$

Looking at the form of ψ_i and p_{ij} it is already clear that it is going to be hopeless to try to find explicit formulae for quantities of interest for this chain. So we're going to look for a diffusion approximation, valid at least for large N.

Now in many interesting cases s, u and v are all $\mathcal{O}(1/N)$ and so we're going to write

$$\alpha = Ns, \quad \mu_1 = 2Nu, \quad \mu_2 = 2Nv.$$

The 2's are convention which lead to the $\frac{1}{2}$ in Bob's $\theta/2$ mutation rate along the branches of his coalescent tree. [Often you'll see the equivalent rescalings with N replaced by $2N$. This corresponds to modelling a population of N diploid individuals as though they were $2N$ haploids.]

So what will the diffusion approximation look like? We will model the *proportion* of individuals of type a rather than absolute numbers - because we want to pass to the infinite population limit. Moreover, as usual - going right back to our discussion of Kingman's coalescent - we'll measure time in units of N generations.

To identify the appropriate diffusion we need to establish $\mathbb{E}[\Delta p]$ and $\mathbb{E}[(\Delta p)^2]$ where Δp is the change in the proportion over a time interval of length Δt as $\Delta t \to 0$. Just as in Example 3.8, our example of a diffusion approximation to a population growth model, we take $\Delta t = \frac{1}{N}$, i.e. one generation. Since the number of type a individuals in the new generation is binomial with N trials and success probability ψ_i, we have

$$\mathbb{E}[\Delta p] = \frac{1}{N}(N\psi_i - i) = \Delta t(N\psi_i - i).$$

Substituting

$$\begin{aligned}
(N\psi_i - i) &= \frac{Ni(1+\frac{\alpha}{N})(1-\frac{\mu_1}{2N})}{N+\frac{\alpha i}{N}} + \frac{(N-i)\frac{\mu_2}{2}}{N+\frac{\alpha i}{N}} - i \\
&= \frac{Ni + \alpha i - \frac{\mu_1}{2}i + \frac{\mu_2}{2}(N-i) - Ni - \frac{\alpha i^2}{N} - \frac{\alpha\mu_1 i}{2N}}{N+\frac{\alpha i}{N}} \\
&= \alpha\left(\frac{i}{N} - \frac{i^2}{N^2}\right) - \frac{\mu_1}{2}\frac{i}{N} + \frac{\mu_2}{2}\left(1 - \frac{i}{N}\right) + \mathcal{O}(\frac{1}{N}) \\
&= \alpha p(1-p) - \frac{\mu_1}{2}p + \frac{\mu_2}{2}(1-p) + \mathcal{O}(\frac{1}{N})
\end{aligned}$$

(where we have used that $0 \le i \le N$).

Similarly,

$$\mathbb{E}[(\Delta p)^2] - (\mathbb{E}[\Delta p])^2 = \frac{1}{N^2}N\psi_i(1-\psi_i)$$

and since

$$\psi_i = \frac{i}{N} + \mathcal{O}(\frac{1}{N})$$

(where again we have used that $0 \le i \le N$) we have

$$\begin{aligned}
var(\Delta p) = \frac{1}{N}\left(\frac{i}{N}(1-\frac{i}{N}) + \mathcal{O}(\frac{1}{N})\right) \\
&= \Delta t p(1-p) + \mathcal{O}(\frac{1}{N})\Delta t.
\end{aligned}$$

As $N \to \infty$ then we see that the limiting diffusion has drift

$$\mu(p) = \alpha p(1-p) - \frac{\mu_1}{2}p + \frac{\mu_2}{2}(1-p),$$

and diffusion coefficient

$$\sigma^2(p) = p(1-p).$$

We are now in a position to calculate exactly as before.

Example 5.1 *Suppose that there is no mutation. If the initial proportion of type a alleles is p, what is the probability that eventually the a-allele* fixes *in the population, that is the proportion of a alleles is absorbed in the boundary point $p = 1$?*

Solution. Using Lemma 4.2 (or rather the work immediately preceding it)

$$P_1(p) = \frac{\int_0^p \exp\left(-2\int_\eta^y \frac{\mu(z)}{\sigma^2(z)}dz\right)dy}{\int_0^1 \exp\left(-2\int_\eta^y \frac{\mu(z)}{\sigma^2(z)}dz\right)dy}$$

where η is some fixed point in $(0,1)$.

In this example $\mu = \alpha p(1-p)$ and $\sigma^2 = p(1-p)$, so

$$\begin{aligned} P_1(p) &= \frac{\int_0^p \exp\left(-2\int_\eta^y \alpha dz\right)dy}{\int_0^1 \exp\left(-2\int_\eta^y \alpha dz\right)dy} \\ &= \begin{cases} \frac{1-\exp(-2\alpha p)}{1-\exp(-2\alpha)} & \text{if } \alpha \neq 0 \\ p & \text{if } \alpha = 0. \end{cases} \end{aligned}$$

□

Now typically we think of the a-allele as arising from a new mutation and we ask about its probability of fixation. Then taking $p = \frac{1}{N}$ we have

$$P_1(\frac{1}{N}) = \begin{cases} \frac{1-\exp(-\frac{2\alpha}{N})}{1-\exp(-2\alpha)} \approx \frac{2\alpha}{N}\frac{1}{(1-\exp(-2\alpha))} & \text{if } \alpha \neq 0 \\ \frac{1}{N} & \text{if } \alpha = o. \end{cases}$$

Recalling that $\alpha = Ns$, we have

$$P_1(\frac{1}{N}) \approx \frac{2s}{(1-\exp(-2Ns))}.$$

Now consider the population growth model of Example 3.8 where we now take $\mu = \alpha p$, $\sigma^2 = p$. For p small, $p(1-p) \approx p$ and so we might hope that this should be a good approximation to the Fisher-Wright diffusion for small gene frequencies. Notice in particular, that if we consider our population growth model only on $(0,1)$ and put an absorbing barrier at 1, then the scale function for the two processes is the same - it depends only on the *ratio* $\frac{2\mu}{\sigma^2}$. They differ only in their speed measure - so only through a local time change. For p small, the ratio of the speed measures is close to one and so the time change makes only little difference. For this reason one often approximates the Fisher-Wright diffusion for small values of p by a population growth model.

If α is large, the new mutant a has a reasonably good chance of fixing in the population. If it does so, then we call the process whereby it increases to fixation a *selective sweep*. A great deal of work has gone into trying to understand selective sweeps and it is still a 'hot topic' in research. As we've said in the early stages, when the allele is at low frequencies, we can approximate its evolution by a population growth model (which turns out to be much easier to study than the full Fisher-Wright diffusion). Once established, its probability of fixation is close to one and as you'll see on the problem sheet it increases extremely rapidly until it is close to fixation - in fact it behaves quasi-deterministically and during this middle phase of the sweep it is often approximated by a deterministic logistic growth model. As it nears fixation, stochastic effects once again become important, but we can approximate $1-p$ by a population 'growth' model with a negative drift - so really a model for population decline.

We now turn to the case of non-zero mutation rates. In that case, the diffusion has a *stationary distribution*. As we calculated from the forward Kolmogorov equation of §3, the density of the stationary distribution will be given by

$$p(y) = \frac{C}{\sigma^2(y)}\exp\left(\int_\eta^y \frac{2\mu(z)}{\sigma^2(z)}dz\right)$$

where the constant is chosen so that $\int_0^1 p(y)dy = 1$. Substituting $\mu = \alpha p(1-p) - \frac{\mu_1}{2}p + \frac{\mu_2}{2}(1-p)$, $\sigma^2 = p(1-p)$ gives

$$p(y) = \text{Const.}e^{2\alpha y}(1-y)^{\mu_1 - 1}y^{\mu_2 - 1}.$$

Armed with the stationary distribution we can calculate various 'summary statistics' for the population. A popular one is the homozygosity.

Definition 5.2 *Take a sample of size two from the population. The probability F that they are of the same allelic type (so both a or both A) is called the* homozygosity.

Example 5.3 *In the selectively neutral case, at stationarity the homozygosity is given by*

$$F = \frac{\mu_1(\mu_1+1) + \mu_2(\mu_2+1)}{(\mu_1+\mu_2)(\mu_1+\mu_2+1)}.$$

In particular, if $\mu_1 = \mu_2 = \mu$ we have $F = \frac{1+\mu}{1+2\mu}$.

Proof. When $s = 0$ the stationary distribution has density

$$p(y) = \text{Const.}y^{\mu_2 - 1}(1-y)^{\mu_1 - 1}.$$

This is a beta-distribution and the constant is

$$\frac{\Gamma(\mu_1+\mu_2)}{\Gamma(\mu_1)\Gamma(\mu_2)}$$

where Γ is the usual Gamma function defined by

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt.$$

(The Gamma function generalises the notion of factorial. In particular, $\Gamma(n+1) = n!$ and $\Gamma(a+1) = a\Gamma(a)$.)

If X is the proportion of a alleles at stationarity, then X has probability density function p and

$$F = \mathbb{E}[X^2 + (1-X)^2].$$

So

$$\begin{aligned} F &= \frac{\Gamma(\mu_1+\mu_2)}{\Gamma(\mu_1)\Gamma(\mu_2)}\int_0^1 \{y^2 + (1-y)^2\}y^{\mu_2-1}(1-y)^{\mu_1-1}dy \\ &= \frac{\Gamma(\mu_1+\mu_2)}{\Gamma(\mu_1)\Gamma(\mu_2)}\int_0^1 \{y^{(\mu_2+2)-1}(1-y)^{\mu_1-1} + y^{\mu_2-1}(1-y)^{(\mu_1+2)-1}\}dy \\ &= \frac{\Gamma(\mu_1+\mu_2)}{\Gamma(\mu_1)\Gamma(\mu_2)}\left\{\frac{\Gamma(\mu_1)\Gamma(\mu_2+2)}{\Gamma(\mu_1+\mu_2+2)} + \frac{\Gamma(\mu_1+2)\Gamma(mu_2)}{\Gamma(\mu_1+\mu_2+2)}\right\} \\ &= \frac{\mu_2(\mu_2+1) + \mu_1(\mu_1+1)}{(\mu_1+\mu_2)(\mu_1+\mu_2+1)}. \end{aligned}$$

□

Notice in this example that if $\mu_1 = \mu_2 = \mu \to \infty$ then the probability of drawing two alleles of the same type tends to $\frac{1}{2}$ - just like flipping a fair coin twice.

If one of the diffusions that we are considering in our genetics world admits a stationary distribution then at stationarity it also has an extremely useful time reversal property: the probability of seeing

a path from x at time 0 to y at time t is the same as that of seeing the 'mirror image' path from y at time $-t$ to x at time 0. This allows us (at stationarity) to say things about past behaviour of the diffusion by studying its future behaviour.

To see the rôle of the stationary distribution in this, let's concentrate for a moment on Markov chains. Then if we are trying to trace backwards in time we are trying to evaluate $\mathbb{P}[X_0 = x|X_t = y]$. Now using Bayes' rule,

$$\mathbb{P}[X_0 = x|X_t = y] = \frac{\mathbb{P}[X_0 = x, X_t = y]}{\mathbb{P}[X_t = y]} = \mathbb{P}[X_t = y|X_0 = x]\frac{\mathbb{P}[X_0 = x]}{\mathbb{P}[X_t = y]}.$$

Now if the chain has reached a stationary distribution π the right hand side can be evaluated - it is $p(t, x, y)\pi(x)/\pi(y)$. To say that the chain is *reversible* is to say that

$$p(t, x, y)\frac{\pi(x)}{\pi(y)} = p(t, y, x), \tag{17}$$

that is, the backwards transitions have the same probabilities as the forwards ones. In the case of Markov chains, equation (17) is called a *detailed balance* equation. We should like something similar to hold in the diffusion setting. To see why it might, we consider the approximating Moran models.

Not all Markov chains are reversible, but the Moran model is just a birth and death process and if a birth and death process admits a stationary distribution then it *is* reversible. Let's check this.

Recall that a birth and death process on $\{0, 1, \ldots, N\}$ is a continuous time Markov process in which if the current state of the process is i, then after the next transition of the chain its state will be either $i - 1$ or $i + 1$. More precisely,

$$\mathbb{P}[X_{t+\delta t} = i + 1|X_t = i] = b_i\delta t + \mathcal{O}(\delta t^2), \quad i = 0, \ldots, N - 1,$$

$$\mathbb{P}[X_{t+\delta t} = i - 1|X_t = i] = d_i\delta t + \mathcal{O}(\delta t^2), \quad i = 1, \ldots, N,$$

and

$$\mathbb{P}[X_{t+\delta t} = j|X_t = i] = \mathcal{O}(\delta t^2), \quad j \notin \{i - 1, i + 1\}.$$

Lemma 5.4 *Suppose that $\{X_t\}_{t\geq 0}$ is a birth and death process. In the notation above, if $\{b_i\}_{i=0}^{N-1}$ and $\{d_i\}_{i=1}^{N}$ are all non-zero, then $\{X_t\}_{t\geq 0}$ has a unique stationary distribution π given by*

$$\pi(i) = \frac{b_0 \cdot b_1 \cdots b_{i-1}}{d_1 \cdot d_2 \cdots d_i}\pi(0),$$

where $\pi(0)$ is determined by $\sum_{i=0}^{N} \pi(i) = 1$. Moreover, $\{X_t\}_{t\geq 0}$ is reversible.

Proof. To check that this is indeed the unique stationary distribution is on the problem sheet. To see that $\{X_t\}_{t\geq 0}$ is reversible, check detailed balance (infinitesimally),

$$p(\delta t, i + 1, i)\frac{\pi(i + 1)}{\pi(i)} = d_{i+1}\delta t\frac{b_i}{d_{i+1}} = b_i\delta t = p(\delta t, i, i + 1).$$

□

Unfortunately, time reversal is not immediately useful for several questions of interest in population genetics because often we are interested in processes for which 0 and/or 1 are absorbing states. In such cases there is not a reversible stationary distribution, but nonetheless we can sometimes make progress.

Example 5.5 *Suppose that in a neutral two-allele model, the a-allele arose as a mutation from an otherwise pure population of A-alleles. If the current frequency of a-alleles is x, how long is it since the mutation first appeared in the population?*

Solution. Since there is just the unique mutation that gave rise to our a population and no further mutation in the model, 0 and 1 are both absorbing states for the frequency of a-alleles. We circumvent this is two ways. Let's think about the Moran model. First, whenever the frequency, p of a-alleles hits zero, we return it to $\frac{1}{N}$ and start a new process. Backwards in time we're still just looking for the time to hit zero. Let's write ϵ for the rate of transitions $0 \mapsto \frac{1}{N}$. Now 1 is also an absorbing state and so we introduce a transition rate $d_N = \epsilon$. The new chain *is* reversible and so we're looking for the forwards (equals backwards) time until the frequency of a-alleles is reduced from x to zero. We then let $\epsilon \to 0$ to recover the age distribution of the a-allele. The only catch is that we know that p *does* hit zero, so we want the *conditional* distribution of this time if we know that 1 is never reached.

There are two things to calculate. First, what is the reversed diffusion and second how do we calculate the *conditioned* diffusion's hitting times?

The time reversal can be shown to be equivalent to reversing with respect to the *speed measure* (so replace π by m in the detailed balance equation). You'll justify this on the problem sheet.

So how do we condition the diffusion to exit $[0,1]$ at 0 rather than 1?

Suppose that the reversed diffusion has generator

$$\mathcal{L}f(x) = \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2} + \mu(x)\frac{df}{dx}$$

and let's write $p(t,x,y)$ for the transition density. It turns out that the process conditioned to hit zero before one is also a diffusion. Let's write $p^*(t,x,y)$ for its transition density and work out what the corresponding generator must be. By Bayes' rule and the Markov property

$$\begin{aligned} p^*(t,x,y) &= \frac{p(t,x,y)\mathbb{P}[\text{exit at 0 started from } y]}{\mathbb{P}[\text{exit at 0 started from } x]} \\ &= p(t,x,y)\left(\frac{S(1)-S(y)}{S(1)-S(x)}\right) \end{aligned}$$

and so

$$\frac{\partial}{\partial t}p^*(t,x,y) = \left(\frac{S(1)-S(y)}{S(1)-S(x)}\right)\frac{\partial}{\partial t}p(t,x,y).$$

We rewrite the right hand side in terms of $\frac{\partial}{\partial x}p^*(t,x,y)$ and $\frac{\partial^2}{\partial x^2}p^*(t,x,y)$. First note that

$$\left(\frac{g}{f}\right)' = \frac{g'}{f} - \frac{gf'}{f^2}$$

(where $'$ denotes differentiation in x) and

$$\left(\frac{g}{f}\right)'' = \frac{g''}{f} - \frac{2g'f'}{f^2} + \frac{2g(f')^2}{f^3} - \frac{gf''}{f^2}.$$

Substituting then

$$\begin{aligned} \mathcal{L}\left(\frac{g}{f}\right) &= \frac{1}{2}\sigma^2\left(\frac{g''}{f} - \frac{2g'f'}{f^2} + \frac{2g(f')^2}{f^3} - \frac{gf''}{f^2}\right) + \mu\left(\frac{g'}{f} - \frac{gf'}{f^2}\right) \\ &= \frac{1}{f}\mathcal{L}g - \frac{g}{f^2}\mathcal{L}f + \frac{\sigma^2 f'}{f}\left\{-\frac{g'}{f} + \frac{gf'}{f^2}\right\} \\ &= \frac{1}{f}\mathcal{L}g - \frac{g}{f^2}\mathcal{L}f - \frac{\sigma^2 f'}{f}\left(\frac{g}{f}\right)'. \end{aligned}$$

Now set g to be $p(t,x,y)$ and

$$f(x) = \frac{S(1)-S(x)}{S(1)-S(y)}.$$

As a function of x, f is just a constant times one minus the scale function and so, in particular, $\mathcal{L}f = 0$ and of course $p^* = \frac{g}{f}$ and $\frac{\partial}{\partial t}p^* = \frac{1}{f}\mathcal{L}g$. Thus

$$\begin{aligned}\frac{\partial}{\partial t}p^*(t,x,y) &= \mathcal{L}\left(\frac{g}{f}\right) + \frac{\sigma^2 f'}{f}\left(\frac{g}{f}\right)' \\ &= \frac{1}{2}\sigma^2\frac{\partial^2}{\partial x^2}p^*(t,x,y) + \mu^*\frac{\partial}{\partial x}p^*(t,x,y)\end{aligned}$$

where

$$\mu^* = \mu + \frac{\sigma^2 f'}{f} = \mu - \sigma^2\frac{S'(x)}{(S(1)-S(x))}.$$

We now have the parameters of our reversed diffusion conditioned to exit $[0,1]$ at 0 and all the ingredients that we need to calculate the expected age of the allele using our results of §4. The final substitutions are left as an exercise. □

6 More than two types

So far we have considered only a very special case in which our population is classified into just two types. The frequencies are then characterised by a one-dimensional diffusion and one dimensional diffusions are, at least in principle, relatively straightforward to study.

More generally, suppose that our population occurs in K different types. We're not going to develop the general theory of multidimensional diffusions, but let's see what happens in a special case. In particular *for the rest of the course all alleles are selectively neutral.*

Our starting point is a K-allele version of the Wright-Fisher model. the population configuration at any time can be described by a vector $(X_1, X_2, \ldots, X_K)$ where X_i is the number of genes of allelic type A_i and we assume that $X_1 + \cdots + X_K = N$. (Although only $K-1$ components are necessary to specify the vector $(X_i)_{i=1}^N$, it is sometimes convenient to retain all K.)

In the simplest case when all the alleles are selectively neutral and there is no mutation, we have

$$\begin{aligned}&\mathbb{P}[Y_i \text{ genes of type } A_i \text{ at } t+1 | X_j \text{ genes of type } A_j \text{ at } t,\ j = 1,\ldots,K] \\ &\qquad = \frac{N!}{Y_1!Y_2!\cdots Y_K!}\psi_1^{Y_1}\psi_2^{Y_2}\cdots\psi_K^{Y_K}\end{aligned}$$

where $\psi_i = \frac{X_i}{N}$ and $\sum_{i=1}^K Y_i = N$ (the probability is zero if this condition is not satisfied).

If we write $p_i = \frac{X_i}{N}$ and δp_i for the change in p_i from one generation to the next, then given $p_1, \ldots, p_{K-1}$,

$$\mathbb{E}[\delta p_i] = 0, \quad var(\delta p_i) = \frac{1}{N}p_i(1-p_i), \quad cov(\delta p_i, \delta p_j) = -\frac{1}{N}p_i p_j (i \neq j).$$

By analogy with what we did in the two-allele case, if we write $f(t, p_1, \ldots, p_{K-1}; p_1', \ldots, p_{K-1}')$ for the joint transition density function we obtain the Kolomogorov backward equation

$$\frac{\partial f}{\partial t} = \frac{1}{2}\sum_i p_i(1-p_i)\frac{\partial^2 f}{\partial p_i^2} - \sum_{i<j} p_i p_j \frac{\partial^2 f}{\partial p_i \partial p_j}.$$

Suppose now that A_i mutates to A_j at the positive rate u_{ij}. Write $\beta_{ij} = 2Nu_{ij}$ (c.f. §5). Now we have

$$\begin{aligned}\mathbb{E}[\delta p_i] &= -p_i \sum_j u_{ij} + \sum_j p_j u_{ji} \\ &= \frac{1}{2}\frac{1}{N} m_i(p_1, \ldots, p_{K-1}),\end{aligned}$$

where

$$m_i(p_1, \ldots, p_{K-1}) = -p_i \sum_j \beta_{ij} + \sum_j p_j \beta_{ji}$$

and the multi-allelic diffusion has generator

$$\frac{1}{2}\sum_i p_i(1-p_i)\frac{\partial^2 f}{\partial p_i^2} - \sum_{i<j} p_i p_j \frac{\partial^2 f}{\partial p_i \partial p_j} + \frac{1}{2} m_i \frac{\partial f}{\partial p_i}.$$

If each $u_{ij} > 0$ for $i \neq j$ then the joint frequency of $A_1, \ldots, A_{K-1}$ has a stationary distribution but no closed form for this has been found in general. Just as in the two-allele case the stationary distribution must satisfy the (forward) equation

$$\begin{aligned}0 = \frac{1}{2}\sum_i \frac{\partial^2}{\partial p_i^2}\left(p_i(1-p_i) f(p_1, \ldots p_{K-1})\right) - \sum_{i<j} \frac{\partial^2}{\partial p_i \partial p_j}\left(p_i p_j f(p_1, \ldots, p_{K-1})\right) \\ - \frac{1}{2}\frac{\partial}{\partial p_i}\left(m_i(p_1, \ldots, p_{K-1}) f(p_1, \ldots, p_{K-1})\right). \quad (18)\end{aligned}$$

In one special case (18) *can* be solved explicitly.

Suppose that $u_{ij} = \frac{u}{K-1}$ so that the total mutation rate per gene is just u and a gene is equally likely to mutate to any of the other types (this is a special case of parent-independent mutation). Then (18) becomes

$$\begin{aligned}0 = \frac{1}{2}\sum_i \frac{\partial^2}{\partial p_i^2}\left(p_i(1-p_i) f(p_1, \ldots p_{K-1})\right) - \sum_{i<j} \frac{\partial^2}{\partial p_i \partial p_j}\left(p_i p_j f(p_1, \ldots, p_{K-1})\right) \\ - \frac{1}{2}\frac{\partial}{\partial p_i}\left(2Nu(1-2p_i) f(p_1, \ldots, p_{K-1})\right), \quad (19)\end{aligned}$$

and this can be solved explicitly to give

$$f(p_1, \ldots, p_{K-1}) = \frac{\Gamma(K\epsilon)}{(\Gamma(\epsilon))^K}(p_1 \cdots p_K)^{\epsilon - 1} \quad (20)$$

where $\epsilon = \frac{2Nu}{K-1}$ and $p_K = 1 - p_1 - \ldots - p_{K-1}$.

Notice that when $K = 2$, (19) becomes

$$0 = -\frac{1}{2}\frac{\partial}{\partial p}\left(\mu(1-2p) f(p)\right) + \frac{1}{2}\frac{\partial^2}{\partial p^2}\left(p(1-p) f(p)\right)$$

where $\mu = 2Nu$ and (20) becomes

$$f(p) = \frac{\Gamma(2\mu)}{(\Gamma(\mu))^2}(p(1-p))^{\mu - 1}$$

which is precisely the solution we found before.

The density (20) is called the *Dirichlet distribution*. It is usual to rearrange it and to consider the sequence of gene frequencies in decreasing order

$$p_{(1)} \geq p_{(2)} \geq \ldots \geq p_{(K)} \geq 0.$$

these frequencies are the *order statistics* of $p_1, \ldots, p_K$ and their joint distribution can be read off directly from (20):

$$f(p_{(1)}, \ldots, p_{(K-1)}) = \frac{K!\Gamma(K\epsilon)}{(\Gamma(\epsilon))^K}(p_{(1)} \cdots p_{(K)})^{\epsilon-1}. \tag{21}$$

The limiting case as $K \to \infty$ is of special interest. Kingman proved that the distribution of the first j order statistics converges as $K \to \infty$ for any j (even though we can't just let $K \to \infty$ in the Dirichlet distribution) and he called the limit the *Poisson-Dirichlet distribution*. It describes the order statistics of the frequencies in the *infinitely many alleles* model in which every mutation leads to a new type.

Let's try to understand why such a limit might exist. Direct manipulation of the Dirichlet distribution is difficult because of the linear dependence between the variables. However, it turns out that it can be represented in terms of *independent* Γ-random variables as follows.

Let $Y_1, \ldots, Y_K$ be independent positive random variables with probability density function

$$g_\epsilon(y) = \frac{y^{\epsilon-1}e^{-y}}{\Gamma(\epsilon)}.$$

Then writing $Y = Y_1 + \cdots + Y_K$, the vector $\mathbf{p}$ with components $p_i = \frac{Y_i}{Y}$ has the Dirichlet distribution and Y has a Γ-distribution with parameter $K\epsilon$. Moreover, $\mathbf{p}$ is *independent* of Y.

The proof of this claim is via a change of variables according to the function $\mathbb{R}^K \to \mathbb{R}^K$ given by

$$(Y_1, \ldots, Y_K) \mapsto (p_1, \ldots, p_{K-1}, Y).$$

Now we can use this representation and in the limit obtain the following representation of the Poisson-Dirichlet distribution.

Take the points of a Poisson process with intensity $\frac{\beta e^{-u}}{u}$, so the number of points in the interval (a, b) is Poisson distributed with mean $\int_a^b \frac{\beta e^{-u}}{u} du$. Then writing $y_{(i)}$ for the *ordered* points and $Y = y_{(1)} + y_{(2)} + \cdots$ we have that Y has a Gamma distribution with parameter β (and recall that $\beta = \lim_{K\to\infty} K\epsilon$) and the points $x_{(i)} = \frac{y_{(i)}}{Y}$ have the Poisson-Dirichlet distribution.

[To see that the convergence really works, we use probability generating functionals. These are a natural extension of the probability generating functions that you learnt about in Mods. So for a random number of randomly positioned points $\{Y_i\}_{i\in I}$ with each $Y_i \in [0, \infty)$ (say) we define the probability generating functional of $\{Y_i\}_{i\in I}$ by

$$G(\xi) = \mathbb{E}[\prod_{i\in I} \xi(Y_i)]$$

for any function $\xi : [0, \infty) \to \mathbb{R}$ for which the expectation exists. (To recover the probability generating function of $|I|$, just choose ξ to be a constant.)

Now choose the Y_i's to be independent Gamma random variables with parameter ϵ and consider the generating functional of $Y_1, \ldots, Y_K$. By independence,

$$G_K(\xi) = \left[\int_0^\infty \xi(u)\frac{u^{\epsilon-1}e^{-u}}{\Gamma(\epsilon)} du\right]^K.$$

Now rewrite the term in square brackets using that

$$\int_0^\infty \frac{u^{\epsilon-1}e^{-u}}{\Gamma(\epsilon)}du = 1 \quad \text{and} \quad \frac{\epsilon}{\Gamma(\epsilon+1)} = \frac{1}{\Gamma(\epsilon)}$$

to obtain

$$G_K(\epsilon) = \left[1 - \epsilon\int_0^\infty (\xi(u)-1)\frac{u^{\epsilon-1}}{\Gamma(\epsilon+1)}e^{-u}du\right]^K$$
$$\to \exp\left(-\beta\int_0^\infty (\xi(u)-1)u^{-1}e^{-u}du\right) \quad \text{as } K\to\infty$$

and this is the probability generating functional of our Poisson points.]

The finite dimensional distributions of the $x_{(i)}$ are complicated, but those of the $y_{(i)}$ are relatively straightforward. The density function of $y_{(i)}$ is

$$\frac{\beta e^{-y}}{y}\frac{[\beta E_1(y)]^{i-1}}{(i-1)!}e^{-\beta E_1(y)}, \quad \text{for } y>0,$$

where $E_1(y) = \int_y^\infty \frac{e^{-u}}{u}du$. Thus, for example,

$$\mathbb{E}[y_{(i)}] = \mathbb{E}[x_{(i)}Y] = \mathbb{E}[x_{(i)}]\mathbb{E}[Y] = \beta\mathbb{E}[x_{(i)}]$$

gives

$$\mathbb{E}[x_{(i)}] = \frac{\beta^{i-1}}{(i-1)!}\int_0^\infty e^{-y}[E_1(y)]^{i-1}e^{-\beta E_1(y)}dy$$

which can be evaluated numerically.

In the Dirichlet distribution with K points, the probability that there are points in $(x_1, x_1+dx_1),\ldots,(x_r,x_r+dx_r)$ is

$$\binom{K}{r}\frac{\Gamma(K\epsilon)}{\Gamma(\epsilon)^r\Gamma((K-r)\epsilon)}(x_1\cdots x_r)^{\epsilon-1}(1-\sum_1^r x_i)^{\epsilon(K-r)-1}dx_1\ldots dx_r$$
$$\to \beta^r(x_1\cdots x_r)^{-1}(1-\sum_1^r x_i)^{\beta-1}dx_1\ldots dx_r \quad \text{as } K\to\infty$$

($\beta = \lim_{K\to\infty} K\epsilon$).

In particular, taking $r=1$, the probability that there is a point in $(x, x+dx)$ for the limiting Poisson-Dirichlet process is $h(x)dx$ where

$$h(x) = \beta x^{-1}(1-x)^{\beta-1}$$

is called the *frequency spectrum* of $\{x_{(i)}\}$.

This allows us to calculate

$$\mathbb{E}[\sum_1^\infty f(x_{(i)})] = \int_0^1 f(x)h(x)dx$$

(provided this is finite). For example, taking $f(x_{(i)}) = x_{(i)}^2$ we calculate the *expected homozygosity*

$$F = \int_0^1 x^2\beta x^{-1}(1-x)^{\beta-1}dx = \frac{1}{1+\beta}.$$

Similarly, the expected number of alleles with frequencies in (a, b) is

$$\mathbb{E}[\sum_{1}^{\infty} \mathbf{1}_{(a,b)}(x_{(i)})] = \int_a^b \beta x^{-1}(1-x)^{\beta-1}dx$$

and so on.

Which allele is oldest?

Let's look at two related questions in this setting:

Question 1. What is the probability that an allele of frequency x in the population is the oldest?

By the reversibility arguments that we used in the two-allele setting (where now we bundle all other types into a single allelic class), this is the same as the probability that it will be the longest lived in the future, which is the probability that our two-allele model will exit $(0, 1)$ at 1, that is x.

So the probability that an allele of frequency x in the population is the oldest is simply x.

Question 2. Let X be the frequency of the oldest allele. What is its probability density function?

$$\begin{aligned}
\mathbb{P}[X \in (x, x+dx)] &= \sum_{j=1}^{\infty} \mathbb{P}[X = x_{(j)}, x_{(j)} \in (x, x+dx)] \\
&= \sum_{j=1}^{\infty} \mathbb{P}[X = x_{(j)} | x_{(j)} = x]\mathbb{P}[x_{(j)} \in (x, x+dx)] \\
&= \sum_{j=1}^{\infty} x\mathbb{P}[x_{(j)} \in (x, x+dx)] \\
&= x\beta x^{-1}(1-x)^{\beta-1}dx \\
&= \beta(1-x)^{\beta-1}dx.
\end{aligned}$$

This can be extended to order the entire population of frequencies by age. This leads to a distribution

$$Z_1, Z_2(1-Z_1), Z_3(1-Z_2)(1-Z_1), \ldots \tag{22}$$

where Z_i are independent identically distributed random variables with density $\beta(1-x)^{\beta-1}$, $0 < x < 1$. For example the expected frequency of the jth oldest allele is

$$\mathbb{E}[Z_j(1-Z_{j-1})\cdots(1-Z_1)] = \frac{1}{1+\beta}\left(\frac{\beta}{1+\beta}\right)^{j-1}, \quad j = 1, 2, \ldots$$

The distribution (22) is called the GEM distribution. 'G' is Bob Griffiths, E and M are Engen and McCloskey.

7 Ewens Sampling formula revisited

Recall the Ewens sampling formula from Bob's lectures. If we take a sample of size n from the infinitely many alleles model, the probability that the sample falls into k distinct allelic types (families) with n_i individuals of type i for each i (where we have imposed some arbitrary order on those types) is

$$\frac{n!\beta^k}{n_1 \cdots n_k \beta(\beta+1)\cdots(\beta+n-1)}.$$

We can also obtain this directly from the Poisson-Dirichlet distribution. Recall that the probability that there are points of the Poisson-Dirichlet process in $(x_1, x_1 + dx_1), \ldots, (x_k, x_k + dx_k)$ is

$$\beta^k (x_1 \cdots x_k)^{-1} (1 - \sum_1^k x_i)^{\beta - 1} dx_1 \ldots dx_k.$$

Now the probability that we see family sizes $n_1, \ldots, n_k$ when we sample from the corresponding partition of $(0, 1)$ (which describes the frequencies in the infinitely many alleles model at stationarity) is the number of ways of assigning the n individuals in our sample to k classes of sizes $n_1, n_2, \ldots, n_k$ times the probability that the first n_1 are from class 1, the next n_2 from class 2 and so on where class i has frequency x_i times the probability that the *are* k points (corresponding to frequencies) in our Poisson-Dirichlet in $(x_1, x_1 + dx_1), \ldots, (x_k, x_k + dx_k)$ integrated over all choices of $x_1, \ldots, x_k$ with $0 \leq x_i \leq 1$ and $\sum_{i=1}^k x_i \leq 1$. That is

$$\begin{aligned}
&\frac{n!}{n_1! \cdots n_k!} \int_{\sum x_i \leq 1} x_1^{n_1} \cdots x_k^{n_k} \beta^k (x_1 \cdots x_k)^{-1} (1 - \sum_1^k x_i)^{\beta - 1} dx_1 \ldots dx_k \\
&\qquad = \frac{n!}{n_1! \cdots n_k!} \beta^k \int x_1^{n_1 - 1} \cdots x_k^{n_k - 1} (1 - \sum_1^k x_i)^{\beta - 1} dx_1 \ldots dx_k \\
&\qquad\qquad = \frac{n!}{n_1! \cdots n_k!} \beta^k \frac{\Gamma(n_1) \cdots \Gamma(n_k) \Gamma(\beta)}{\Gamma(n + \beta)} \\
&\qquad\qquad\qquad = \frac{n! \beta^k}{n_1 \cdots n_k} \frac{1}{\beta(\beta + 1) \cdots (\beta + n - 1)}
\end{aligned}$$

as required. □