1.8 The site frequency spectrum

The simplest statistic for a sample under the infinitely many sites mutation model is the number of segregating sites, whose distribution we discussed in §1.5, but one can also ask for more detailed information.

Definition 1.23 (Site frequency spectrum) *For a sample of size k under the infinitely many sites mutation model, write $M_j(k)$ for the number of sites at which exactly j individuals carry a mutation. The vector $(M_1(k), M_2(k), \ldots, M_k(k))$ is called the* site frequency spectrum *of the sample.*

PICTURE

Lemma 1.24 *If the genealogy of the sample is determined by the Kingman coalescent, then under the infinitely many sites model we have*

$$\mathbb{E}[M_j(k)] = \frac{\theta}{j}. \tag{5}$$

Note: the θ here is as before, so mutations occur at rate $\theta/2$ along each ancestral lineage.

Proof of Lemma 1.24

We use the duality between the Kingman coalescent and the Moran model. Suppose that a mutation arose at time $-t$ (that is t before the present) and denote individuals in our sample carrying that mutation as type a. For the dual Moran model (with population size k), we think of the sample as the whole population and so the Moran population has nothing to do with the population from which we are sampling. From the point of view of the Moran model, the probability that we see j type a individuals in the sample is the probability that a mutation arising on a single individual at time zero is carried by j individuals at time t later. We write X_t for the number of type a individuals at time t and $p(t, i, j) = \mathbb{P}[X_t = j | X_0 = i]$. In this notation, the probability that at time 0 there are exactly j type a individuals in the sample is $p(t, 1, j)$.

Since each mutation occurs at a different point on the genome and mutations occur at rate $\theta/2$ per individual (and the population size is k), the expected *total* number of sites at which we see a mutation carried by exactly j individuals is then just

$$\mathbb{E}[M_j(k)] = \int_0^\infty k\frac{\theta}{2}p(t, 1, j)dt. \tag{6}$$

Now $G(i, j) \equiv \int_0^\infty p(t, i, j)dt$ is just the expected total time that the process $\{X_t\}_{t\geq 0}$ spends in site j if it started from i and our next task is to calculate this.

Note that if $X_s = i$, then it moves to a new value at rate $i(k - i)$ (which is just the number of the $\binom{k}{2}$ ways of sampling a pair from the population in which the two individuals sampled are of different types) and when it does move, it is equally likely to move to $i - 1$ or $i + 1$. Let

$$T_i = \inf\{t > 0 : X_t = i\}$$

denote the first hitting time of site i. Then since 0 is a trap for the process we have

$$G(1,j) = \mathbb{P}[T_j < T_0 | \, X_0 = 1] \cdot G(j,j).$$

Now, because it is just a timechange of a simple random walk,for $0 \leq i \leq j$,

$$\mathbb{P}[T_0 < T_j | \, X_0 = i] = \frac{j-i}{j},$$

and similarly, for $j \leq l \leq k$,

$$\mathbb{P}[T_k < T_j | \, X_0 = l] = \frac{l-j}{k-j}.$$

Thus, partitioning on whether the first jump out of j is to $j-1$ or to $j+1$, we find that if it is currently at j, the probability that this is the *last* visit that X_t makes to j is

$$\rho = \frac{1}{2}\frac{1}{j} + \frac{1}{2}\frac{1}{k-j} = \frac{1}{2}\frac{k}{j(k-j)}.$$

In other words, if we start from j, the number of visits to j (including the current one) before either the allele is fixed in the population or lost is geometric with parameter ρ. Each visit lasts an exponentially distributed time with mean $\frac{1}{j(k-j)}$. Thus

$$G(1,j) = \frac{1}{j}G(j,j) = \frac{1}{j}\frac{1}{\rho}\frac{1}{j(k-j)} = \frac{2}{kj}.$$

Substituting into (6) completes the proof. □

1.9 The lookdown process

The consistency of the k-coalescents for different values of $k \in \mathbb{N}$ allowed us to recover all of them as projections of a single stochastic process, Kingman's coalescent. Since genealogical trees for the Moran model are precisely governed by the Kingman coalescent, it is reasonable to hope that we can also construct Moran models corresponding to different population sizes as projections of a single stochastic process. This is at the heart of the powerful Donnelly & Kurtz *lookdown process*.

To see how it works, we exploit the connection with the Kingman coalescent. Suppose that the population at the present time is labelled $\{1, 2, \ldots, N\}$. Recall that the full description of the Kingman coalescent is as a process taking values among the set of equivalence relations on $\{1, 2, \ldots, N\}$, with each ancestral lineage corresponding to a single equivalence class. Now suppose that we label each equivalence class by its smallest element. If blocks with labels $i < j$ coalesce, then after the coalescence the new block is necessarily labelled i. In our graphical representation of the Moran model, this just dictates the direction of the arrow corresponding to that coalescence event; it will always be the individual with the smaller label that gave birth. Backwards in time, our process is equivalent to one

in which, as before, at the points of a rate one Poisson process $\pi_{(i,j)}$ arrows are drawn joining the labels i and j, but now the arrows are always in the same direction (upwards with our convention). The genealogies are still determined by the Kingman coalescent, we have simply chosen a convenient labelling, and so in particular they are precisely those of the Moran model. But what about forwards in time? What we saw backwards in time was that choosing the direction of the arrows corresponded to choosing a particular labelling of the population. If the distribution of the population is *exchangeable*, that is it doesn't depend on the labelling, then forwards in time too we should not have changed the distribution in our population. Our next task will be to check this, but first we need a formal definition.

Definition 1.25 (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', with $\zeta_i(t)$ denoting the allelic type of the individual at level i at time t. 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 it selects a level uniformly at random from $\{1, 2, \ldots, k-1\}$ and adopts the current type of the individual at that level. The levels of the individuals involved in the event do not change.*

Remark 1.26 *Because of our convention over the interpretation of arrows, it is not at all clear from the above why one should call this the* lookdown *process. The explanation is that at rate $(k-1)$ the kth individual looks* down *to a level chosen uniformly at random from those below and adopts the type of the individual at that level.*

To see that the lookdown process and the Moran model produce the same distribution of types in the population, provided we start from an exchangeable initial condition, we examine their infinitesimal generators. Recall the definition of the generator of a continuous time Markov process.

Definition 1.27 (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 $\{X_t\}_{t\geq 0}$.*

If we know $\mathcal{L}$, then we 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 $\{X_t\}_{t\geq 0}$.

We suppose that the types of individuals are sampled from some type space E. The Moran model for a population of size N is then simply a continuous time Markov chain on E^N and its infinitesimal generator, K_N, evaluated on a function $f : E^N \to \mathbb{R}$, is given by

$$K_N f(x_1, x_2, \ldots, x_N) = \sum_{i=1}^{N} A_i f(x_1, x_2, \ldots, x_N) + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\Phi_{ij} f(x_1, \ldots, x_N) - f(x_1, \ldots, x_N)\right], \quad (7)$$

where $\Phi_{ij} f(x_1, \ldots, x_N)$ is the function obtained from f by replacing x_j by x_i. The operator A_i is the generator of the mutation process, A, acting on the ith coordinate. (Recall that in the Moran model mutation was superposed as a Markov process along lineages.)

The generator of the N-particle lookdown process, L_N is given by

$$L_N f(x_1, x_2, \ldots, x_N) = \sum_{i=1}^{N} A_i f(x_1, x_2, \ldots, x_N) + \sum_{1 \le i < j \le N} \left[\Phi_{ij} f(x_1, x_2 \ldots, x_N) - f(x_1, x_2 \ldots, x_N) \right]. \tag{8}$$

Assuming that we start both processes from the same exchangeable initial condition, we should like to show that, $(\zeta_1(t), \zeta_2(t), \ldots, \zeta_N(t))$ and the types under the original Moran process which we denote $(Z_1(t), Z_2(t), \ldots, Z_N(t))$ have the same distribution for each fixed $t > 0$, even though the processes are manifestly different.

Following Dawson (1993), we check that the generators of the two processes agree on symmetric functions. Observe first that any symmetric function, f, satisfies

$$f(x_1, x_2, \ldots, x_N) = \frac{1}{N!} \sum_{\pi} f\left(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi_N}\right),$$

where the sum is over all permutations of $\{1, 2, \ldots, N\}$. Substituting this expression for f into equation (8), we recover precisely equation (7). In other words, the generators of $(\zeta_1, \zeta_2, \ldots, \zeta_N)$ and $(Z_1, Z_2, \ldots, Z_N)$ agree on symmetric functions as required. (We are implicitly assuming uniqueness of the distribution on symmetric functions corresponding to this generator. It follows from duality with the N-coalescent, but we don't allow that to detain us here.)

The key observation now is that our Nth lookdown process is simply the first N levels of the $(N+k)$th lookdown process for any $k \ge 1$. The *infinite* lookdown process can then be constructed as a projective limit.

Theorem 1.28 (Donnelly & Kurtz 1996) *There is an infinite exchangeable particle system $\{W_i, i \in \mathbb{N}\}$ such that for each N,*

$$(W_1, W_2, \ldots, W_N) \stackrel{\mathcal{D}}{=} (\zeta_1, \zeta_2, \ldots, \zeta_N),$$

where $\zeta_1, \zeta_2, \ldots, \zeta_N$ is the N-particle lookdown process.

Remark 1.29 *In fact more is true. It is known that the sequence of empirical measures $\frac{1}{N} \sum_{i=1}^{N} \delta_{Z_i(t)}$ converges to a* Fleming-Viot superprocess *as $N \to \infty$. Donnelly & Kurtz also show that*

$$Y = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \delta_{W_i},$$

is a Fleming-Viot superprocess. Rather than introduce the general Fleming-Viot superprocess, which takes its values among probability measures on the type space E, in §1.10 we shall consider what this limit looks like in the special case when E is a two-point set representing two alleles a and A in which case it is enough to specify the evolution of the proportion of type a individuals in the population.

Since the genealogy of a sample of size k from the Moran model is a k-coalescent and since we've seen that the genealogy of the first k levels in the lookdown process is also a k-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 k 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.

1.10 A more simplistic limit

Rather than discussing general Fleming-Viot superprocesses (which would allow us to consider essentially arbitrary type spaces) we now turn to identifying the limiting model for allele frequencies when our population is subdivided into just two types which, as usual, we label a and A. Just as in our discussion of the rescaled Wright-Fisher model, we consider the proportion, p_t, of individuals of type a at time t.

The only possible mutations are between the two types. We suppose that each type a individual mutates to type A at rate ν_1 and each type A individual mutates to type a at rate ν_2. Recall that for the Moran model we are already in the timescale of the Kingman coalescent and so we should think of $\nu_i = N\mu_i$ where μ_1 and μ_2 are the true mutation rates.

Remark 1.30 *The idea that we can mutate backwards and forwards between types may seem at odds with our discussion of mutations in §1.3. Models of this type were introduced long before biologists knew about and had access to DNA sequences. Classically one might imagine a small number of alleles defined through phenotype, for example colour. In modern terms one can justify the model by pooling sequences into classes according to the corresponding phenotype.*

The generator for the Moran model for a population of size N is then

$$\mathcal{L}_N f(p) = \binom{N}{2} p(1-p)\left(f(p+\frac{1}{N}) - f(p)\right) + \binom{N}{2} p(1-p)\left(f(p-\frac{1}{N}) - f(p)\right) + N\nu_1 p\left(f(p-\frac{1}{N}) - f(p)\right) + N\nu_2(1-p)\left(f(p+\frac{1}{N}) - f(p)\right).$$

To see this, note that the reproduction events in 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, if the current proportion of a alleles is p, then

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

and there is no change with probability $1 - 2p(1-p)$. The chance that we see a reproduction event 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 $\mathcal{O}((\delta t)^2)$. For mutation events, at total rate $Np\nu_1$, one of the Np type a individuals will mutate to type A, resulting in a reduction of p by $1/N$ and at total rate $N(1-p)\nu_2$ one of the $N(1-p)$ type A individuals will mutate to type a. Putting all this together gives that for $f:[0,1]\to\mathbb{R}$ and $p=\frac{i}{N}$ for some $i\in\{0,1,\ldots,N\}$

$$\mathcal{L}_N f(p) = \binom{N}{2}p(1-p)\left(f(p+\frac{1}{N})-f(p)\right)+\binom{N}{2}p(1-p)\left(f(p-\frac{1}{N})-f(p)\right) \\ + Np\nu_1\left(f(p+\frac{1}{N})-f(p)\right)+Np\nu_2\left(f(p-\frac{1}{N})-f(p)\right).$$

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

$$\begin{aligned}\mathcal{L}_N f(p) &= \binom{N}{2}p(1-p)\Big(f(p)+\frac{1}{N}f'(p)+\frac{1}{2N^2}f''(p)+\mathcal{O}(\frac{1}{N^3})-f(p)\Big)\\ &\quad+\binom{N}{2}p(1-p)\Big(f(p)-\frac{1}{N}f'(p)+\frac{1}{2N^2}f''(p)+\mathcal{O}(\frac{1}{N^3})-f(p)\Big)\\ &\quad-Np\nu_1\Big(f(p)+\frac{1}{N}f'(p)+\mathcal{O}(\frac{1}{N^2}\Big)+N(1-p)\nu_2\Big(f(p)-\frac{1}{N}f'(p)+\mathcal{O}(\frac{1}{N^2}\Big)\\ &= \frac{1}{2}p(1-p)f''(p)+((1-p)\nu_2-p\nu_1)f'(p)+\mathcal{O}(\frac{1}{N}).\end{aligned}$$

So as $N\to\infty$, $\mathcal{L}_N\to\mathcal{L}$ where

$$\mathcal{L}f(p)=\frac{d}{dt}\mathbb{E}[f(p_t)|p_0=p]\Big|_{t=0}=\frac{1}{2}p(1-p)f''(p)+(\nu_2-(\nu_1+\nu_2)p)f'(p). \tag{9}$$

In particular, if we set $\nu_1=\nu_2=0$ we obtain

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

which is exactly the generator that we obtained in the large population limit from our Wright-Fisher model. It is not hard to extend the work that we did there to include mutations and recover the full generator (9).

What we have written down is the generator of a one-dimensional diffusion. We should like to be able to use the convergence of generators that we have verified to justify using the corresponding one-dimensional diffusion as an approximation for the Moran, Wright-Fisher and Cannings models (on suitable timescales).

Theorem 1.31 *Let E be a metric space. Suppose that for each $N\in\mathbb{N}$, $\{X_t^{(N)}\}_{t\geq 0}$ is an E-valued Markov processes with generator $\mathcal{L}^N$ and that X is an E-valued Markov process with generator $\mathcal{L}$. If, for every $f\in\mathcal{D}(\mathcal{L})$,*

$$\lim_{N\to\infty}\mathcal{L}^N f(x)=\mathcal{L}f(x),\quad \text{uniformly } \textit{for } x\in E,$$

then the finite-dimensional distributions of $X^{(N)}$ converge to those of X. That is, for every finite set of times $0 \leq t_1 < t_2 < \cdots < t_n$,

$$(X^{(N)}(t_1), \ldots, X^{(N)}(t_n)) \xrightarrow{d} (X(t_1), \ldots, X(t_n)) \quad \text{as } N \to \infty.$$

In fact we have used slightly more than this as our Wright-Fisher model was in discrete time. For that Ethier & Kurtz, Chapter 1, Theorem 6.5 is exactly what we need.

Remark 1.32 *This sort of convergence is enough to justify using our limiting Wright-Fisher diffusion to approximate things like time to fixation and fixation probabilities. However, if we are really interested in the genealogies of populations, then we need more. For our Moran models, the Donnelly-Kurtz lookdown construction gave us a much stronger result. In general we must be careful. It is possible to arrive at the same diffusion for allele frequencies from many different individual based models for our population, and it is* not *always the case that the genealogies converge to the same limit.*

Before we can exploit Theorem 1.31 we need to know that there is a Markov process with generator (9) and that we can actually calculate quantities of interest for it. Happily both are true.

1.11 Diffusions

In this section we are going to remind ourselves of some useful facts about one-dimensional diffusions. We start in a fairly general setting.

Definition 1.33 (One-dimensional diffusion) *A one-dimensional diffusion process $\{X_t\}_{t\geq 0}$ is a strong Markov process on $\mathbb{R}$ which traces out a continuous path as time evolves.*

At any instant in time, X_t is a continuous random variable but also any realisation of $\{X_t\}_{t\geq 0}$ is a continuous function of time. Its range need not be the whole of $\mathbb{R}$ and indeed for the most part we'll be interested in diffusions on $(0, 1)$. For the time being let us take the state space to be an interval (a, b) (possibly infinite). The generator of the diffusion takes the form

$$\mathcal{L}f(x) = \frac{1}{2}\sigma^2(x)\frac{d^2 f}{dx^2}(x) + \mu(x)\frac{df}{dx}(x). \tag{10}$$

Evidently for this to be defined f must be twice continuously differentiable on (a, b). Depending on the behaviour of the diffusion close to the boundaries of its domain, f may also have to satisfy boundary conditions at a and b. We'll specify these precisely in Theorem 1.45, but for now assume that if we apply the generator to a function then it is in the domain. To avoid pathologies, we make the following assumptions:

1. For any compact interval $I \subset (a, b)$, there exists $\epsilon > 0$ such that $\sigma^2(x) > \epsilon$ for all $x \in I$,
2. the coefficients $\mu(x)$ and $\sigma^2(x)$ are continuous functions of $x \in (a, b)$.

Note that (crucially for applications in genetics) we *do* allow $\sigma^2(x)$ to vanish at the boundary points $\{a, b\}$. This is (more than) enough to guarantee that the diffusion has a transition density function, denoted $p(t, x, y)$ (see Knight 1981 Theorem 4.3.5 for a more general result).

Definition 1.34 *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 $A \subseteq \mathbb{R}$.

Let us write $\Delta_h X(t) = X_{t+h} - X_t$, then taking $f_1(x) = x$ in the generator (and using the Markov property) we see that

$$\mathcal{L}f_1(X_t) = \lim_{h\downarrow 0} \frac{1}{h}\mathbb{E}[\Delta_h X(t) |\, X_t] = \mu(X_t)$$

and so

$$\mathbb{E}[\Delta_h X(t) |\, X_t] = h\mu(X_t) + o(h) \qquad \text{as } h \downarrow 0. \tag{11}$$

Now observe that we can write $(X_{t+h} - X_t)^2 = X_{t+h}^2 - X_t^2 - 2X_t(X_{t+h} - X_t)$ and so, taking $f_2(x) = x^2$,

$$\mathcal{L}f_2(X_t) - 2X_t\mathcal{L}f_1(X_t) = \lim_{h\downarrow 0} \frac{1}{h}\mathbb{E}[(\Delta_h X(t))^2 \big|\, X_t] = \sigma^2(X_t)$$

which yields

$$\mathbb{E}[(\Delta_h X(t))^2 \big|\, X_t] = h\sigma^2(X_t) + o(h) \qquad \text{as } h \downarrow 0. \tag{12}$$

This motivates the standard terminology.

Definition 1.35 (Infinitesimal drift and variance) *The coefficients* $\mu(x)$ *and* $\sigma^2(x)$ *are called the (infinitesimal)* drift *and* variance *of the diffusion* $\{X_t\}_{t\geq 0}$.

In fact if a strong Markov process $\{X_t\}_{t\geq 0}$ is càdlàg (that is its paths are right continuous with left limits) and satisfies (11), (12) and the additional condition

$$\lim_{h\downarrow 0} \frac{1}{h}\mathbb{E}[|\Delta_h X(t)|^p |\, X_t = x] = 0 \qquad \text{for some } p > 2$$

where the convergence is uniform in (x, t) on compact subsets of $(a, b) \times \mathbb{R}_+$, then $\{X_t\}_{t\geq 0}$ is necessarily a diffusion (see Karlin & Taylor, §15.1, Lemma 1.1).

The canonical example of a one-dimensional diffusion is one-dimensional Brownian motion which has generator

$$\mathcal{L}_B f(x) = \frac{1}{2}\frac{d^2 f}{dx^2}(x)$$

and transition density function

$$p(t,x,y) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(x-y)^2}{2t}\right).$$

Brownian motion can be thought of as a building block from which other one-dimensional diffusions are constructed. One approach is to observe that the diffusion corresponding to the generator $\mathcal{L}$ of equation (10) can be expressed as the solution of a stochastic differential equation driven by Brownian motion (with appropriate boundary conditions)

$$dX_t = \mu(X_t)dt + \sigma(X_t)dB_t.$$

Remark 1.36 (Mathematical drift versus genetic drift) *We have already encountered the Wright-Fisher diffusion several times, corresponding to the solution of the stochastic differential equation*

$$dp_t = (\nu_2(1-p_t) - \nu_1 p_t)dt + \sqrt{p_t(1-p_t)}dB_t.$$

It is an unfortunate accident of history, that the standard terminology for the stochastic term (driven by Brownian motion) is genetic drift*, whereas to a mathematician it is the deterministic mutation term that corresponds to drift.*

1.11.1 Speed and scale

Our approach to constructing one-dimensional diffusions from Brownian motion will not be via stochastic differential equations, but rather through the theory of speed and scale. A nice feature of one dimensional diffusions is that many quantities can be calculated explicitly. This is because (except at certain singular points which will only ever be at a or b under our conditions) all one-dimensional diffusions can be transformed into Brownian motion first by a change of space variable (through the so-called scale function) and then a timechange (through what is known as the speed measure).

To see how this works, we first investigate what happens to a diffusion when we change the timescale. Suppose that a diffusion $\{Z_t\}_{t\geq 0}$ has generator $\mathcal{L}_Z$. We define a new process $\{Y_t\}_{t\geq 0}$ by $Y_t = Z_{\tau(t)}$ where

$$\tau(t) = \int_0^t \beta(Y_s)ds,$$

for some function $\beta(x)$ which we assume to be bounded, continuous and strictly positive. So if $Y_0 = Z_0$, then the increment of Y_t over an infinitesimal time interval $(0, dt)$ is that of Z_t over the interval $(0, d\tau(t)) = (0, \beta(Y_0)dt)$. In our previous notation,

$$\mathbb{E}[\Delta_h Y(0)|Y(0)] = \beta(Y_0)h\mu_Z(Z_0) + o(h) = \beta(Y_0)\mu_Z(Y_0)h + o(h),$$

and

$$\mathbb{E}[(\Delta_h Y)^2|Y_0] = \beta(Y_0)h\sigma_Z^2(Z_0) + o(h) = \beta(Y_0)\sigma^2(Y_0)h + o(h).$$

In other words,

$$\mathcal{L}_Y f(x) = \beta(x)\mathcal{L}_Z f(x).$$

We are now in a position to understand speed and scale. Let $\{X_t\}_{t\geq 0}$ be governed by the generator (10). Suppose now that $S(x)$ is a strictly increasing function on (a,b) and consider the new process $Z_t = S(X_t)$. Then the generator $\mathcal{L}_Z$ of Z can be calculated as

$$\begin{aligned}
\mathcal{L}_Z f(x) &= \frac{d}{dt}\mathbb{E}[\, f(Z_t)|\, Z_0 = x]\Big|_{t=0} \\
&= \frac{d}{dt}\mathbb{E}[\, f(S(X_t))|\, S(X_0) = x]\Big|_{t=0} \\
&= \frac{1}{2}\sigma^2(S^{-1}(x))\frac{d^2}{dx^2}(f(x) + \mu(S^{-1}(x))\frac{d}{dx}f(x) \\
&= \frac{1}{2}\sigma^2(S^{-1}(x))\left\{(S'(x))^2\frac{d^2 f}{dx^2}(x) + S''(x)\frac{df}{dx}(x)\right\} + \mu(S^{-1}(x))S'(x)\frac{df}{dx}(x) \\
&= \frac{1}{2}\sigma^2(S^{-1}(x))S'(S^{-1}(x))^2\frac{d^2}{dx^2}f(x) + \mathcal{L}S(x)\frac{d}{dx}f(x). \qquad (13)
\end{aligned}$$

Now if we can find a strictly increasing function S that satisfies $\mathcal{L}S \equiv 0$, then the drift term (in the mathematical sense) in (13) will vanish and so Z_t will just be a time change of Brownian motion on the interval $(S(a), S(b))$. Such an S is provided by the scale function of the diffusion.

Definition 1.37 (Scale function) *For a diffusion X_t on (a,b) with drift μ and variance 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).

The scale change resulted in a timechanged Brownian motion. The change of time required to transform this into standard Brownian motion is dictated by the speed measure.

Definition 1.38 (Speed measure) *The function $m(\xi) = \frac{1}{\sigma^2(\xi)S'(\xi)}$ is the* density of the speed measure *or just the* speed density *of the process X_t. We write*

$$M(x) = \int_{x_0}^{x} m(\xi)d\xi.$$

Remark 1.39 *The function m plays the rôle of β before. Notice that*

$$\int_{x_0}^{x} m(\xi)d\xi = \int_{S(x_0)}^{S(x)} m(S^{-1}(y))\frac{1}{S'(S^{-1}(y))}dy = \int_{S(x_0)}^{S(x)} \frac{1}{\sigma^2(S^{-1}(y))\big(S'(S^{-1}(y))\big)^2}dy.$$

The additional $S'(y)$ in the generator (13) has been absorbed since our time change is applied to a diffusion on $(S(a), S(b))$.

In summary, we have the following.

Lemma 1.40 *Denoting the scale function and the speed measure by S and M respectively we have*

$$\mathcal{L}f = \frac{1}{2}\frac{1}{dM/dS}\frac{d^2f}{dS^2} = \frac{1}{2}\frac{d}{dM}\left(\frac{df}{dS}\right).$$

Proof

$$\begin{aligned}
\frac{1}{2}\frac{d}{dM}\left(\frac{df}{dS}\right) &= \frac{1}{2}\frac{1}{dM/dx}\frac{d}{dx}\left(\frac{1}{dS/dx}\frac{df}{dx}\right) \\
&= \frac{1}{2}\sigma^2(x)S'(x)\frac{d}{dx}\left(\frac{1}{S'(x)}\frac{df}{dx}\right) \\
&= \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2} - \frac{1}{2}\sigma^2(x)S'(x)\frac{S''(x)}{(S'(x))^2}\frac{df}{dx} \\
&= \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2} + \mu(x)\frac{df}{dx}
\end{aligned}$$

(since S solves $\mathcal{L}S = 0$) as required. □

1.11.2 Hitting probabilities and Feller's boundary classification

Before going further, let's see how we might apply this. Suppose that a diffusion process on $(0, 1)$ represents the frequency of an allele, a say, in a population and that zero and one are traps for the process. One question that we should like to answer is "What is the probability that the a-allele is eventually lost from the population?" In other words, what is the probability that the diffusion hits zero before one? To prove a general result we need first to be able to answer this question for Brownian motion.

Lemma 1.41 *Let $\{B_t\}_{t\geq 0}$ be standard Brownian motion on the line. For each $y \in \mathbb{R}$, let T_y denote the random time at which it hits y for the* first *time. Then for $a < x < b$,*

$$\mathbb{P}[T_a < T_b |\, B_0 = x] = \frac{b-x}{b-a}.$$

Sketch of Proof

Let $u(x) = \mathbb{P}[T_a < T_b |\, B_0 = x]$ and choose h small enough that $\mathbb{P}[T_a \wedge T_b < h | B_0 = x] = o(h)$. We suppose that u is twice differentiable, then

$$\begin{aligned}
u(x) &= \mathbb{E}[\,u(B_h)|\, B_0 = x] + o(h) \\
&= \mathbb{E}[u(x) + (B_h - x)u'(x) + \frac{1}{2}(B_h - x)^2u''(x)] + o(h) \\
&= u(x) + \frac{1}{2}hu''(x) + o(h).
\end{aligned}$$

Subtracting $u(x)$ from each side, dividing by h and letting h tend to zero, we obtain $u''(x) = 0$. We also have the boundary conditions $u(a) = 1$ and $u(b) = 0$. This is easily solved to give

$$u(x) = \frac{b-x}{b-a}$$

as required. □ Of course this reflects the corresponding result for simple random walk that we used in the proof of Lemma **??**. In general we can reduce the corresponding question for $\{X_t\}_{t\geq 0}$ to solution of the equation $\mathcal{L}u(x) = 0$ with $u(a) = 1$ and $u(b) = 0$, but in fact we have already done all the work we need. We have the following result.

Lemma 1.42 (Hitting probabilities) *Let $\{X_t\}_{t\geq 0}$ be a one-dimensional diffusion on (a, b) with infinitesimal drift $\mu(x)$ and variance $\sigma^2(x)$ satisfying the conditions above. If $a < a_0 < x < b_0 < b$ then*

$$\mathbb{P}[T_{a_0} < T_{b_0} | X_0 = x] = \frac{S(b_0) - S(X_0)}{S(b_0) - S(a_0)},$$

where S is the scale function for the diffusion.

Remark 1.43 *Notice that η cancels in the ratio and x_0 in the difference, so that this ratio is well-defined.*

Proof

Evidently it is enough to consider the corresponding hitting probabilities for the process $Z_t = S(X_t)$, where S is the scale function. The process Z_t is a time changed Brownian motion, but since we only care about *where* not *when* the process exits the interval $(S(a_0), S(b_0))$, then we need only determine the hitting probabilities for Brownian motion and the result follows immediately from Lemma 1.41. □

Before continuing to calculate quantities of interest, we fill in a gap left earlier when we failed to completely specify the domain of the generators of our one-dimensional diffusions. Whether or not functions in the domain must satisfy boundary conditions at a and b is determined by the nature of those boundaries from the perspective of the diffusion. More precisely, we have the following classification.

Definition 1.44 (Feller's boundary classification) *Define*

$$u(x) = \int_{x_0}^{x} M dS, \qquad v(x) = \int_{x_0}^{x} S dM.$$

The boundary b is said to be

regular	*if*	$u(b) < \infty$	*and*	$v(b) < \infty$
exit	*if*	$u(b) < \infty$	*and*	$v(b) = \infty$
entrance	*if*	$u(b) = \infty$	*and*	$v(b) < \infty$
natural	*if* $u(b) = \infty$	*and*	$v(b) = \infty$	

with symmetric definitions at a.

Regular and exit boundaries are said to be accessible *while entrance and natural boundaries are called* inaccessible.

Theorem 1.45 *The domain of the generator (10) is continuous functions f on $[a,b]$ which are twice continuously differentiable on the interior and for which*

1. *if a and b are inaccessible there are no further conditions,*
2. *if b (resp. a) is an exit boundary, then*

$$\lim_{x\to b} \mathcal{L}f(x) = 0 \qquad \left(resp.\ \lim_{x\to a} \mathcal{L}f(x) = 0\right).$$

If b (resp. a) is a regular boundary, then for each $q \in [0,1]$ we get a different process by restricting f in the domain to satisfy

$$q \lim_{x\to b} \mathcal{L}f(x) = (1-q) \lim_{x\to b} S'(x)f'(x) \qquad \left(resp.\ q \lim_{x\to a} \mathcal{L}f(x) = -(1-q) \lim_{x\to a} S'(x)f'(x)\right).$$

For a more careful discussion see Ethier & Kurtz (1986), Chapter 8.

1.11.3 Green's functions

Lemma 1.42 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 $\{X_t\}_{t\geq 0}$ is a Markov process,

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

and so for $a < p < b$

$$w(p) = \mathbb{E}[\int_0^h g(X_s)ds|X_0 = p] + \mathbb{E}[w(X_h)|X_0 = p]. \tag{14}$$

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

$$\mathbb{E}[\int_0^h g(X_s)ds|X_0 = p] = hg(p) + \mathcal{O}(h^2). \tag{15}$$

Now subtract $w(p)$ from both sides of (14), divide by h and let $h \downarrow 0$ to obtain

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

Let us now turn to solving this equation. Using Lemma 1.40 we have

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

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 $\int_a^b = \int_p^b + \int_a^p$.

Theorem 1.46 *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 1.37 and $m(\xi) = \frac{1}{\sigma^2(\xi)S'(\xi)}$, the density of the speed measure.

Definition 1.47 *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.*

Example 1.48 *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, $S(x) = x$ (up to an arbitrary additive constant). What about $\mathbb{E}[T^*]$?

Using Theorem 1.46 with $g = 1$ we have

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

□

In our Moran model, at least if the population is large, then we expect that if the current proportion of a alleles is p, the time until either the a allele or the A allele is fixed in the population has mean approximately

$$-2\left\{p\log p + (1-p)\log(1-p)\right\}. \tag{17}$$

In fact by conditioning on whether the proportion of a-alleles increases or decreases at the first reproduction event, one obtains a recurrence relation for the *number of jumps* until the process first hits either zero or one. This recurrence relation can be solved explicitly and since jumps ocur at independent exponentially distributed times with mean $1/\binom{N}{2}$, it is easy to verify that (17) is indeed a good approximation. For the Wright-Fisher model, in its original timescale, there is no explicit expression for the expected time to fixation, $t(p)$. However, since changes in p over a single generation are typically small, one can expand $t(p)$ in a Taylor series, in just the way we did to derive equation (1) and thus verify that for a large population,

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

This is readily solved to give

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

just as predicted by our diffusion approximation.

1.11.4 Stationary distributions and reversibility

Before moving on to models in which a gene is allowed to have more than two alleles, we consider one last quantity for our one-dimensonial diffusions. First a general definition.

Definition 1.49 (Stationary distribution) *Let $\{X_t\}_{t\geq 0}$ be a Markov process on the space E. A* stationary distribution *for $\{X_t\}_{t\geq 0}$ is a probability distribution ψ on E such that if X_0 has distribution ψ, then X_t has distribution ψ for all $t \geq 0$.*

In particular this definition tells us that if ψ is a stationary distribution for $\{X_t\}_{t\geq 0}$, then

$$\frac{d}{dt}\mathbb{E}\left[f(X_t) \middle| X_0 \sim \psi \right] = 0,$$

where we have used $X_0 \sim \psi$ to indicate that X_0 is distributed according to ψ. In other words

$$\frac{d}{dt} \int_E \mathbb{E}\left[f(X_t) \middle| X_0 = x \right] \psi(dx) = 0.$$

Evaluating the time derivative at $t = 0$ gives

$$\int_E \mathcal{L} f(x) \psi(dx) = 0.$$

Sometimes this allows us to find an explicit expression for $\psi(dx)$. Let $\{X_t\}_{t\geq 0}$ be a one-dimensional diffusion on (a, b) with generator given by (10). We're going to suppose that there is a stationary distribution which is absolutely continuous with respect to Lebesgue measure. Let us abuse notation a

little by using $\psi(x)$ to denote the density of $\psi(dx)$ on (a,b). Then, integrating by parts, we have that for all $f \in \mathcal{D}(\mathcal{L})$,

$$\begin{aligned} 0 &= \int_a^b \left\{ \frac{1}{2}\sigma^2(x)\frac{d^2f}{dx^2}(x) + \mu(x)\frac{df}{dx}(x) \right\} \psi(x)dx \\ &= \int_a^b f(x) \left\{ \frac{1}{2}\frac{d^2}{dx^2}\left(\sigma^2(x)\psi(x)\right) - \frac{d}{dx}\left(\mu(x)\psi(x)\right) \right\} dx + \text{boundary terms.} \end{aligned}$$

This equation must hold for all f in the domain of $\mathcal{L}$ and so, in particular,

$$\frac{1}{2}\frac{d^2}{dx^2}\left(\sigma^2(x)\psi(x)\right) - \frac{d}{dx}\left(\mu(x)\psi(x)\right) = 0.$$

Integrating once gives

$$\frac{1}{2}\frac{d}{dx}\left(\sigma^2(x)\psi(x)\right) - \left(\mu(x)\psi(x)\right) = C_1,$$

for some constant C_1 and then using $S'(x)$ as in integrating factor we obtain

$$\frac{d}{dy}\left(S'(y)\sigma^2(y)\psi(y)\right) = C_1 S'(y),$$

from which

$$\psi(x) = C_1\frac{S(x)}{S'(x)\sigma^2(x)} + C_2\frac{1}{S'(x)\sigma^2(x)} = m(x)\left[C_1 S(x) + C_2\right].$$

If can arrange constants so that $\psi \geq 0$ and

$$\int_a^b \psi(\xi)d\xi = 1$$

then the stationary density exists and equals ψ. In particular, if $\int_a^b m(y)dy < \infty$, then

$$\psi(x) = \frac{m(x)}{\int_a^b m(y)dy}$$

is a stationary measure for the diffusion.

Example 1.50 *Recall the generator of the Wright-Fisher diffusion with mutation,*

$$\mathcal{L}f(x) = \frac{1}{2}x(1-x)\frac{d^2f}{dx^2} + \Big(\nu_2(1-x) - \nu_2 x\Big)\frac{df}{dx}.$$