1.3 Forward Kolmogorov equation

Let us again start with the Master equation, for a system where the states can be ordered along a line, such as the previous examples with population size $n = 0, 1, 2 \cdots, N$. We start again with a general Master equation

$$\frac{dp_n}{dt} = -\sum_{m \neq n} R_{mn} p_n + \sum_{m \neq n} R_{nm} p_m \,. \tag{1.28}$$

In many relevant circumstances, the number of states is large and the probability varies smoothly from one site to the next. In such cases, it is reasonable to replace the discrete index n with a continuous variable x, the probabilities $p_n(t)$ with a probability density $p(x,t)$, and the rates R_{mn} with a rate function $R(x', x)$. The rate function R depends on two variables denoting the start and end positions along the line. We are free to redefine the two arguments of this function, and it is useful to reparametrize it as $R(x' - x, x)$ indicating the rate at which, starting from the position x, a transition is made to a position $\Delta = x' - x$ away. As in the case of mutations, there is usually a preference for changes that are *local*, i.e. the rates decay rapidly when the separation $x' - x$ becomes large.

These transformations and relabelings,

$$n \to x, \;\; p_n \to p(x,t), \;\; R_{mn} \to R(x' - x, x)\,, \tag{1.29}$$

enable us to transform Eq. (1.28) to the continuous integral equation

$$\frac{\partial}{\partial t} p(x,t) = -\int^{*} dx' R(x' - x, x) p(x,t) + \int^{*} dx' R(x - x', x') p(x', t). \tag{1.30}$$

Note, however, that the sum in Eq. (1.28) excluded the term $m = n$. To treat this properly in the continuum limit, we can focus on an interval y around any point x, and the change in probability due to incoming flux from $x - y$ and the outgoing flux to $x + y$, leading to

$$\frac{\partial}{\partial t} p(x,t) = \int^{*} dy \left[R(y, x - y) p(x - y) - R(y, x) p(x) \right]. \tag{1.31}$$

We now make a Taylor expansion the first term in the square bracket, *but only with respect to the location of the incoming flux*, treating the argument pertaining to the separation of the two points as fixed, i.e.

$$R(y, x - y) p(x - y) = R(y, x) p(x) - y \frac{\partial}{\partial x} \left(R(y, x) p(x) \right) + \frac{y^2}{2} \frac{\partial^2}{\partial x^2} \left(R(y, x) p(x) \right) + \cdots \,. \tag{1.32}$$

While formally correct, the above expansion is useful only in cases where typical values of y are small (only almost *local* transitions occur). If we keep terms up to the second order, Eq. (1.31) can be rewritten as

$$\frac{\partial}{\partial t} p(x,t) = -\int dy \, y \frac{\partial}{\partial x} (R(y, x) p(x)) + \frac{1}{2} \int dy \, y^2 \frac{\partial^2}{\partial x^2} (R(y, x) p(x)). \tag{1.33}$$

The integrals over y can be taken inside the derivatives with respect to x,

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}\left[p(x)\left(\int dy\, yR(y,x)\right)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[p(x)\left(\int dy\, y^2R(y,x)\right)\right] , \tag{1.34}$$

after which we obtain

$$\boxed{\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[v(x)\ p(x,t)\right] + \frac{\partial^2}{\partial x^2}\left[D(x)p(x,t)\right].} \tag{1.35}$$

We have introduced

$$v(x) \equiv \int dy\, yR(y,x) = \frac{\langle \Delta(x)\rangle}{\Delta t} , \tag{1.36}$$

and

$$D(x) \equiv \frac{1}{2}\int dy\, y^2R(y,x) = \frac{1}{2}\frac{\langle \Delta(x)^2\rangle}{\Delta t} . \tag{1.37}$$

Equation (1.35) is a prototypical description of *drift* and *diffusion* which appears in many contexts. The *drift* term $v(x)$ expresses the rate (velocity) with which the position changes from x due to the transition rates. Given the probabilistic nature of the process, there are variations in the rate of change of position captured by the position dependent *diffusion* coefficient $D(x)$. The drift–diffusion equation is known as the *forward Kolmogorov* equation in the context of populations. As a description of random walks it appeared earlier in physics literature as the *Fokker–Planck* equation.

In the context of population genetics, it is convenient to introduce the variable $x = n/N$, such that in the continuum limit $x \in [0, 1]$. The rates in Eq. (1.22) change n by ±1, and hence

$$v(x) = \frac{\langle \Delta n\rangle}{N} = \frac{R_{n+1,n}\times 1 + R_{n-1,n}\times(-1)}{N} = \frac{1}{N}\left[\mu_1(N-n) - \mu_2 n\right] = \mu_1(1-x) - \mu_2 x , \tag{1.38}$$

while

$$D(x) = \frac{\langle \Delta n^2\rangle}{2N^2} = \frac{R_{n+1,n} + R_{n-1,n}}{2N^2} = \frac{1}{2N^2}\left[\mu_1(N-n) + \mu_2 n\right] = \frac{\mu_1(1-x)+\mu_2 x}{2N} . \tag{1.39}$$

1.3.1 Binomial selection

Consider a population with two forms of an allele, say A_1 and A_2 corresponding to blue or brown eye colors. The probability for a spontaneous mutation to occur that changes the allele for eye color is extremely small, and effectively $\mu_1 = \mu_2 = 0$ in Eq. (1.23). Yet the proportions of the two alleles in the population does change from generation to generation. One reason is that some individuals do not reproduce and leave no descendants, while others reproduce many times and have multiple descendants. This is itself a stochastic process and the major source of rapid changes in allele proportions. In principle this effect also leads to variations in population size. In practice, and to simplify computations, it is typically assumed that the size of the population is fixed.

In the model of *binomial selection*, the process or reproduction from one generation to the next is assumed to be as follows: Let us assume that in a population of N alleles, $N_1 = n$ are A_1, and $N - n$ are A_2. The population at the next generation may have m individuals with allele A_1, and the probability for such a transition is

$$\Pi_{mn} = \left(\frac{n}{N}\right)^m \left(1 - \frac{n}{N}\right)^{N-m} \left(\begin{array}{c} N \\ m \end{array}\right) . \tag{1.40}$$

This probability is like reaching into a bag with n balls of blue color and $N-m$ balls of brown color, recording the color of the ball and throwing it back. After repeating the process N times, the probability that the blue color is recorded m times is given by the above binomial distribution. (The probability of getting a blue ball in each trial is simply n/N, and $1-n/N$ for brown.) Clearly some balls can be picked up multiple times (multiple descendants), while some balls are never picked (no offspring).

Regarding R_{mn} as the probability to obtain random variable m, given initial n, it is easy to deduce from standard properties of the binomial distribution that

$$\langle m \rangle = N \times \frac{n}{N} = n \,, \quad \text{i.e} \quad \langle (m-n) \rangle = 0 \,, \tag{1.41}$$

while

$$\left\langle m^2 \right\rangle_c = \left\langle (m-n)^2 \right\rangle = N \times \frac{n}{N} \left(1 - \frac{n}{N}\right) . \tag{1.42}$$

We can construct a continuum evolution equation by setting $x = n/N \in [0, 1]$, and replacing $p(n, t+1) - p(n, t) \approx dp(x)/dt$, where t is measured in number of generations. Clearly, from Eq. (1.41), there is no drift

$$v(x) = \langle (m-n) \rangle = 0 \,, \tag{1.43}$$

while the diffusion coefficient is given by

$$D_{\text{haploid}}(x) = \frac{1}{2N^2} \left\langle (m-n)^2 \right\rangle = \frac{1}{2N} x(1-x) \,. \tag{1.44}$$

A light variant of binomial selection is also applicable to mating of diploid organism. For two alleles, there are three genotypes of A_1A_1, A_1A_2, and A_2A_2 in proportions of x_{11}, x_{12}, and x_{22} respectively. To mimic a mating event, pick one allele of one individual, another allele from a second individual. Set aside the resulting offspring and return the parents to the initial pool. Repeat the process N times to construct the new generation. For each offspring the probability of selecting allele A_1 is $x_{11} + x_{12}/2$, while allele A_2 is selected with probability $x_{22} + x_{12}/2$. If the initial population is in *Hardy–Weinberg equilibrium*, the relative genotype frequencies are related to the proportions of the two alleles simply by $x_{11} = x^2$, $x_{12} = 2x_1x_2$, and $x_{22} = x_2^2$. The mating process is thus again equivalent to the process we considered earlier for haploids, with A_1 and A_2 chosen with probabilities of x and $1 - x$ respectively. Since, in a diploid population of N individuals, the number of alleles is $2N$, the previous result is simply modified to

$$D_{\text{diploid}}(x) = \frac{1}{4N} x(1-x) \,. \tag{1.45}$$

1.3.2 Chemical analog & Selection

Through the reactions in Eq. (1.25), we introduced a chemical reaction that mimicks a mutating population. Consider a system where a reaction between molecules A and B can lead to two outcomes:[2]

$$\mathrm{A}+\mathrm{B}\rightharpoondown_{c}\mathrm{A}+\mathrm{A}\quad\text{or}\quad\mathrm{A}+\mathrm{B}\rightharpoondown^{d}\mathrm{B}+\mathrm{B}\,, \tag{1.46}$$

at rates c and d. In a "mean-field" approximation the number of A molecules changes as

$$\frac{dN_A}{dt}=(c-d)N_AN_B=(c-d)N_A(N-N_A)\,. \tag{1.47}$$

Equation (1.47) predicts steady states $N_A^*=0$ for $c<d$, $N_A^*=N$ for $c>d$, while any composition is permitted for the symmetric case of $c=d$. As we shall demonstrate, fluctuations modify the latter conclusion.

As before, let us denote $N_A=n$, $N_B=N-N_A$, and follow the change in composition after a single reaction. The number of A particles may change by ±1 with rates

$$R_{n,n+1}=d(n+1)(N-n-1),\quad\text{and}\quad R_{n,n-1}=c(n-1)(N-n+1)\,, \tag{1.48}$$

where the product is over the number of possible pairs of A-B particles that can participate in the reaction. The diagonal terms are again obtained from the normalization condition in Eq. (1.14) resulting in the Master equation

$$\frac{dp(n,t)}{dt}=d(n{+}1)(N{-}n{-}1)p(n{+}1)+c(n{-}1)(N{-}n{+}1)p(n{-}1)-dn(N{-}n)p(n)-cn(N{-}n)p(n)\,, \tag{1.49}$$

for $0<n<N$, and with boundary terms

$$\frac{dp(0,t)}{dt}=d(N-1)p(1),\quad\text{and}\quad\frac{dp(N,t)}{dt}=c(N-1)p(N-1)\,. \tag{1.50}$$

When the number N is large, it is reasonable to take the continuum limit and construct a Kolmogorov equation for the fraction $x=n/N\in[0,1]$. The rates in Eq. (1.48) change n by ±1, and hence

$$\begin{aligned}v(x)&=\frac{\langle\Delta n\rangle}{N}=\frac{R_{n+1,n}-R_{n-1,n}}{N}=\frac{1}{N}\left[cn(N-n)-dn(N-n)\right]\\&=N(c-d)x(1-x)\,,\end{aligned} \tag{1.51}$$

while

$$\begin{aligned}D(x)&=\frac{\langle\Delta n^2\rangle}{2N^2}=\frac{R_{n+1,n}+R_{n-1,n}}{2N^2}=\frac{1}{2N^2}\left[cn(N-n)+dn(N-n)\right]\\&=\frac{c+d}{2}x(1-x)\,.\end{aligned} \tag{1.52}$$

[2] In a sense these reactions mimic the mating process in which the offspring of a *heterozygote* (a diploid organism with different alleles A_1 and A_2) and a *homozygote* (say with two copies of allele A_1) may be either heterozygote (A_1A_2) or homozygote (A_1A_1).

Comparison with Eqs.(1.43) and Eq. (1.45) indicates that the above reaction has the same behavior as binomial selection provided that $c = d = 1/(4N)$. Indeed the superficial difference in factor of N between the two cases is because in the latter we followed the reactions one at a time (at rate $c = d$), while in the former we computed the transition probabilities after a whole generation (N steps of reproduction and removal). The selection process characterized by Eq.(1.40) treats the two alleles as completely equivalent. In reality one allele may provide some advantage to individuals carrying it. If so, there should be a *selection* process by which individuals with this allele are more likely to reproduce, on average increasing their population in the next generation. This would then cause a drift in the appropriate Kolmogorov equation. The population genetics perspective on selection will be covered in detail by Professor Mirny. It turns out that this prescription is mathematically equivalent to the binary reaction of Eq. (1.46) with $c \neq d$. In future lectures, selection is quantified by a parameter s, which is related to c and d by

$$c = \frac{1}{4N}(1+s) \quad \text{and} \quad d = \frac{1}{4N}(1-s)\ . \tag{1.53}$$

In the following, we shall employ the nomenclature of population genetics, such that

$$v(x) = \frac{s}{2}x(1-x)\,, \quad \text{and} \quad D(x) = \frac{1}{4N}x(1-x)\,. \tag{1.54}$$

1.3.3 Steady states

While it is usually hard to solve the Kolmogorov equation as a function of time, it is relatively easy to find the steady state solution to which the population settles after a long time. Let us denote the steady-state probability distribution by $p^*(x)$, which by definition must satisfy

$$\frac{\partial p^*(x)}{\partial t} = 0. \tag{1.55}$$

Therefore, setting the right-hand side of Eq. (1.35) to zero, we get

$$-\frac{\partial}{\partial x}\left[v(x)p^*(x)\right] + \frac{\partial^2}{\partial x^2}\left[D(x)p^*(x)\right] = 0. \tag{1.56}$$

The most general solution admits steady states in which there is an overall current and the integral over x of the last equation leads to a constant flow in probability. It is not clear how such a circumstance may arise in the context of population genetics, and we shall therefore focus on circumstances where there is no probability current, such that

$$-v(x)p^*(x) + \frac{\partial}{\partial x}(D(x)p^*(x)) = 0. \tag{1.57}$$

We can easily rearrange this equation to

$$\frac{1}{D(x)p^*}\frac{\partial}{\partial x}(D(x)p^*(x)) = \frac{\partial}{\partial x}\ln\left(D(x)p^*(x)\right) = \frac{v(x)}{D(x)}\,. \tag{1.58}$$

This equation can be integrated to

$$\ln D(x)p^*(x) = \int^x dx' \frac{v(x')}{D(x')} + \text{constant}, \tag{1.59}$$

such that

$$p^*(x) \propto \frac{1}{D(x)} \exp\left[\int^x \frac{v(x')}{D(x')}\right] , \tag{1.60}$$

with the proportionality constant set by boundary conditions.

Let us examine the case of the dynamics of a fixed population, including mutations, and reproduction with selection. Adding the contributions in Eqs. (1.38), (1.39) and (1.54), we have

$$v(x) = \frac{s}{2}x(1-x) + \mu_1(1-x) - \mu_2 x\,, \tag{1.61}$$

while

$$D(x) = \frac{1}{4N}x(1-x) + \frac{\mu_1(1-x)+\mu_2 x}{2N} \approx \frac{1}{4N}x(1-x)\,. \tag{1.62}$$

The last approximation of ignoring the contribution from mutations to diffusion is common to population genetics and we shall follow it here without further justification. It enables a closed form solution to the steady state, as

$$\begin{aligned}
\log D(x)p^*(x) &= \int^x dx' \frac{v(x')}{D(x')} \\
&= 4N \int^x dx' \left[\frac{\mu_1}{x'} - \frac{\mu_2}{1-x'} + \frac{s}{2}\right] \\
&= 4N \left[\mu_1 \ln x + \mu_2 \ln(1-x) + \frac{s}{2}x\right] + \text{constant},
\end{aligned}$$

resulting in

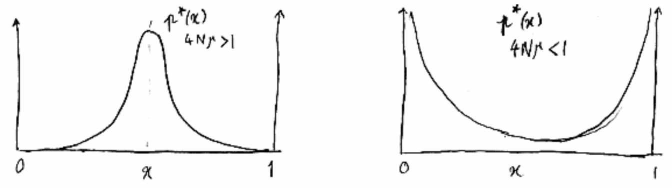
$$p^*(x) \propto \frac{1}{x(1-x)} \cdot x^{4N\mu_1} \cdot (1-x)^{4N\mu_2} \cdot e^{2Nsx}\,. \tag{1.63}$$

In the special case of no selection, $s = 0$ and (for convenience) $\mu_1 = \mu_2 = \mu$, the steady-state solution (1.63) simplifies to

$$p^*(x) \propto [x(1-x)]^{4N\mu-1}\,. \tag{1.64}$$

The shape of the solution is determined by the parameter $4N\mu$. If $4N\mu > 1$, then the distribution has a peak at $x = 1/2$ and diminishes to the sides. On the other hand, if the population is small and $4N\mu < 1$, then $p^*(x)$ has peaks at either extreme—a situation where *genetic drift* is dominant.

* No selection ($\bar{w} = 1$); neutral mutations ($\mu_1 = \mu_2 = \mu$)


$p^*(x)$
$4N\mu > 1$
x
$p^*(x)$
$4N\mu < 1$
x