Chapter 6

Infinitely many types models

Figure 6.1: Kimura

6.1 Introduction

6.1.1 Motivation

We have considered particle systems above with finitely many types. In the 1970's with the advent of electrophoresis and molecular biology, new models were needed in which the number of types were not fixed. In many cases the number of types can be random and new types can be introduced at random times. Several models began to appear at that time involving infinitely many types, for example the *ladder* or *stepwise mutation model* of Ohta and Kimura (1973) [492] (which could model for example continuous characteristics). Another model was one in which no attempt to model the structure of types was made but in which new types can be introduced (leading to the *infinitely many alleles* model) (Kimura and Crow (1964) [398]). In this model we take $[0, 1]$ as the type space. Then when a new type is needed we can choose a type in $[0, 1]$ by sampling from the uniform distribution on $[0, 1]$. The *infinitely many sites model* introduced by Kimura in 1969 provides an idealization of the genome viewed as a sequence of nucleotides (A,T,C,G). These processes now form the basis for molecular population genetics.

More generally, such infinitely many type models provide the possibility of coding information at a number of levels and provide a powerful tool for the study of complex systems. For example we can code historical information, genealogical information, and information about the random environment that has been visited. In addition it allows for individuals with internal structure described by an internal state space and state transition dynamics.

6.1.2 Plan

The objective of this chapter is to construct two basic infinitely-many-type processes, formulated as measure-valued processes, by taking the projective limit of the finite type Feller CSB and Wright-Fisher diffusions. These are the *Jirina measure-valued branching process* and *infinitely many alleles model* of Crow and Kimura. The latter has played a central role in population genetics. We will establish a relation between the invariant measures of these two processes that allows us to obtain the basic properties of the *Poisson-Dirichlet* distribution and the *Griffiths, Engen and McCloskey (GEM) representation.*

We begin by considering the diffusion limit of a measure-valued generalization of the Wright-Fisher Markov chain in the setting of semigroup theory. This process, the *Fleming-Viot process*, includes the infinitely many alleles model as a special case. In the next Chapter we reformulate these processes in terms of measure-valued martingale problems and develop techniques for working with more general classes of measure-valued processes including the class of superprocesses and the class of Fleming-Viot processes with selection, mutation and recombination.

6.2 Measure-valued Wright-Fisher Markov chain

We now consider a Wright-Fisher model of a population of N individuals in which the space of types is a separable metric space E. The process is then a Markov chain $\{p_n^N\}_{n\in\mathbb{N}}$ with state space

$$\mathcal{P}^N(E) = \left\{ \mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i},\ \{x_1, \ldots, x_N\} \in E \right\} \subset \mathcal{P}(E).$$

In this case the mutation process is a Markov chain on E with probability transition function $P(x, dy)$ giving the type distribution of the offspring of a type x parent if mutation occurs. Let $V > 0$ be a measurable function on E with $V(x)$ interpreted as the (haploid) fitness of a type x individual.

Then as in the finitely many type case X_n^N is a Markov chain in $\mathcal{P}^N(E)$ with one step transition function $P(\mu, d\mu^{\text{next}})$. This is obtained by noting that X_{n+1}^N

is a random probability measure on $\mathcal{P}^N(E)$ given by:

$$X_{n+1}^N = \frac{1}{N}\sum_{i=1}^{N}\delta_{y_i} \tag{6.1}$$

where $y_1, \ldots, y_N$ are i.i.d. $\mu_n^*(dy)$ where

$$\mu_n^*(dy) = \int_E \left(\frac{V(x)X_n^N(dx)}{\int V(x)X_n^N(dx)}\right) P(x, dy),$$

that is, as before selection first and then mutation and sampling.

Example 6.1 *Infinitely many alleles model [398]. $E = [0, 1]$ and*

$$P(x, dz) = (1 - m)\delta_x(dz) + m\int_0^1 \delta_y(dz)\lambda(dy) \tag{6.2}$$

where λ is Lebesgue measure on $[0, 1]$.

Example 6.2 *The infinitely many sites model was introduced by Kimura [400], [401]. (See also Ethier and Griffiths (1987) [233])*

Infinitely many sites model. $E = [0, 1]^{\mathbb{Z}_+}$

$$P(\mathbf{x}, dy) = (1 - m)\delta_{\mathbf{x}}(dy) + m\int_0^1 \delta_{\{\xi,\mathbf{x})}\lambda(d\xi)$$

Here we interpret ξ as the locus on the genome where the last mutation occurred.

The number of segregating sites is the number of homologous DNA positions that differ in a sample of m sequences. They are used to investigate phylogenetic relationships. The location of polymorphisms within humans are also used to determine the potential differences in reactions of individuals to medical treatments.

See Section 8.3.3 for the analysis of segregating sites.

Example 6.3 *Ladder model of Ohta and Kimura (1973) [492]. This stepwise-mutation model was introduced to describe the distribution of allelic types distinguishable as signed electrical charges in gel electrophoresis experiments.*
Here $E = \mathbb{Z}$ and

$$P(x, dy) = (1 - m)\delta_x(\cdot) + \frac{m}{2}\delta_{x+1} + \frac{m}{2}\delta_{x-1} \tag{6.3}$$

6.3 The neutral Fleming-Viot process with mutation generator A

The infinitely many alleles diffusion of Crow and Kimura can be studied as an infinite dimensional diffusion (i.e. countably many types) (see Ethier and Kurtz

(1981) [232]) in which a mutation always leads to a new type. However it is advantageous to reformulate it as a measure-valued process. This process was introduced by Fleming and Viot in 1979 [255]. We will show that it arises as the diffusion limit of the measure-valued Wright-Fisher model.

We will now we derive the Fleming-Viot process under some simplifying assumptions using semigroup methods. The general case will be dealt with below in the martingale problem setting.

Assumptions

- Let E be a compact metric space.
- $V \equiv 1$, that is, we omit the selection effect.
- We consider a mutation process given by a Feller process on E with generator $(D(A), A)$ and semigroup $\{S_t : t \geq 0\}$ on $C(E)$. A will be called the mutation operator for the Fleming-Viot process.

  We assume that $D(A)$ contains an algebra $D_0(A)$ that separates points and $S_t : D_0(A) \to D_0(A)$. Then linear combinations of functions in $D_0(A)$ form an algebra of functions separating points and therefore is dense in $C(E)$ and therefore measure-determining. We also assume that A arises as the limit of a sequence of mutation Markov chains on E with transition kernels $\{P_N(.,.)\}_{N\in\mathbb{N}}$, that is for $f \in D(A)$,

  (6.4) $N(\langle f, P_N\rangle - f) \to Af$ as $N \to \infty$

  uniformly on E.

The state space for the Fleming-Viot (FV) process is $\mathcal{P}(E)$, the set of Borel probability measures on E with the topology of weak convergence. For $f \in C(E)$, $\mu \in \mathcal{P}$ we denote $\langle f, \mu\rangle = \int f d\mu$.

We will now obtain the neutral FV process as the limit of neutral (i.e. $V \equiv 1$) Wright-Fisher Markov chains in the diffusion time scale,

(6.5) $p^N(t) = X^N_{\lfloor Nt\rfloor} \in \mathcal{P}(E)$

where X^N_n is defined by (6.1) with mutation transition functions $P^N(x, dy)$.

In order to identify the limiting generator for a $\mathcal{P}(E)$-valued diffusion we need a measure-determining family of test functions. Consider the algebra $\mathcal{D}$ of nice functions on $\mathcal{P}(E)$ containing the functions:

(6.6) $F(\mu) = \langle f_1, \mu\rangle \ldots \langle f_n, \mu\rangle$

with $n \geq 1$ and $f_1, \ldots, f_n \in D(A)$. This algebra of functions is measure-determining in $\mathcal{P}(\mathcal{P}(E))$.

Notation 6.4 *For $F \in \mathcal{D}$, $x \in E$ we define*

$$\frac{\partial F(\mu)}{\partial \mu(x)} = \lim_{\varepsilon \to 0} \frac{F(\mu + \varepsilon \delta_x) - F(\mu)}{\varepsilon}|_{\varepsilon=0} = \frac{\partial F(\mu + \varepsilon \delta_x)}{\partial \varepsilon}|_{\varepsilon=0}$$
$$\frac{\partial^2 F(\mu)}{\partial \mu(x) \partial \mu(y)} = \frac{\partial^2 F(\mu + \varepsilon_1 \delta_x + \varepsilon_2 \delta_y)}{\partial \varepsilon_1 \partial \varepsilon_2}|_{\varepsilon_1=\varepsilon_2=0}$$

Proposition 6.5 *Let p_n^N denote the measure-valued Wright-Fisher Markov chain (6.1) under the above assumptions. Then $p^N(t) = X^N_{\lfloor Nt \rfloor} \Rightarrow p_t$ where $\{p_t\}_{t \geq 0}$ is a $\mathcal{P}(E)$-valued Markov process with generator*

$$\begin{aligned}
&GF(\mu) \\
&= \sum_{1 \leq i < j \leq n} \left(\langle f_i f_j, \mu \rangle - \langle f_i, \mu \rangle \langle f_j, \mu \rangle \right) \prod_{\ell : \ell \neq i,j} \langle f_\ell, \mu \rangle + \sum_i \langle A f_i, \mu \rangle \prod_{\ell : \ell \neq i} \langle f_\ell, \mu \rangle \\
&= \frac{1}{2} \left[\int \frac{\partial^2 F(\mu)}{\partial \mu(x) \partial \mu(y)} \delta_x(dy) \mu(dx) - \int \frac{\partial^2 F(\mu)}{\partial \mu(x) \partial \mu(y)} \mu(dx) \mu(dy) \right] \\
&+ \int A \frac{\partial F(\mu)}{\mu(x)} \mu(dx).
\end{aligned}$$

for all $F \in \mathcal{D}$.

Proof. Here we follow the Ethier-Kurtz semigroup approach. Using the Kurtz semigroup convergence theorem ([225], Chap. 1, Theorem 6.5, Proposition 3.7 and Chap. 4. Theorem 2.5 -see Appendix I Theorems 18.1, 18.2). Using these results it suffices to show that for $F \in \mathcal{D}$,

$$\lim_{N \to \infty} N E_\mu[F(p^N_{\frac{1}{N}}) - F(\mu))] = \lim_{N \to \infty} N E_\mu[F(X_1^N) - F(\mu)] = GF(\mu) \tag{6.7}$$

uniformly in $\mu \in \mathcal{P}(E)$.

First note that for $f_1, \ldots, f_n \in C(E)$

$$E_\mu(F(X_1^N)) = E_\mu[\langle f_1, X_1^N \rangle \ldots \langle f_n, X_1^N \rangle], \quad F(\mu) = \langle f_1, \mu \rangle \ldots \langle f_n, \mu \rangle$$

where $X_1^N = \frac{1}{N} \sum_{i=1}^N \delta_{Y_i}$ and $Y_1, \ldots, Y_N$ are i.i.d. μP^N. Hence

$$\begin{aligned}
E_\mu[\langle f_1, X_1^N \rangle \ldots \langle f_n, X_1^N \rangle] &= \frac{1}{N^n} E \left[\sum_{i=1}^N f_1(Y_i) \cdots \sum_{i=1}^N f_n(Y_i) \right] \\
&= \sum_{k=1}^n \frac{N^{[k]}}{N^n} \sum_{\beta \in \pi(n,k)} \prod_{j=1}^k \left\langle \prod_{i \in \beta_j} f_i, \mu P^N \right\rangle \\
&= \sum_{k=1}^n \frac{N^{[k]}}{N^n} \sum_{\beta \in \pi(n,k)} \prod_{j=1}^k \left\langle \left\langle \prod_{i \in \beta_j} f_i, P^N \right\rangle, \mu \right\rangle
\end{aligned}$$

where $N^{[k]} = \frac{N!}{(N-k)!}$, $\pi(n,k)$ is the set of partitions β of $\{1,\dots,n\}$ into k nonempty subsets $\beta_1,\dots,\beta_k$, labelled so that $\min\beta_1 < \cdots < \min\beta_k$.

Only the terms involving $k = n, n-1$ contribute in the limit. To see this note that we can choose n different Y_i's in $N(N-1)\dots(N-n+1) = N^n - \frac{n(n-1)}{2}N^{n-1} + O(N^{n-2})$ ways and $n-1$ different Y_i's in $N(N-1)\dots(N-n+2) = N^{n-1} - O(N^{n-2})$ ways. For $k = n-2$ we can choose k different Y_i's is $O(N^{n-2})$ ways, etc.

$$
\begin{aligned}
&NE_\mu[F(X_1^N) - F(\mu)] \\
&= N\left\{\frac{N^{[n]}}{N^n}\prod_{j=1}^n\langle f_j,\mu P_N\rangle + \frac{N^{[n-1]}}{N^n}\sum_{1\le i<j\le n}\langle f_if_j,\mu P_N\rangle\prod_{\ell:\ell\neq i,j}\langle f_\ell,\mu P_N\rangle\right. \\
&\left. + O(N^{-2}) - \prod_{j=1}^n\langle f_j,\mu\rangle\right\} \\
&= N\left\{\left(1-\frac{n(n-1)}{2N}\right)\prod_{j=1}^n\langle f_j,\mu P_N\rangle\right. \\
&\left. + \frac{1}{N}\sum_{1\le i<j\le n}\langle f_if_j,\mu P_N\rangle\prod_{\ell\neq i,j}\langle f_\ell,\mu P_N\rangle - \prod_{j=1}^n\langle f_j,\mu\rangle\right\} + O(\frac{1}{N})
\end{aligned}
$$

Note that $\lim_{N\to\infty}\langle f,\mu P_N\rangle = \langle f,\mu\rangle$. Now let $b_j = \langle f_j,\mu\rangle$ and $a_j = \langle f_j,\mu P_N\rangle$ and recall that

$$\lim_{N\to\infty} N(\langle f,\mu P_N\rangle - \langle f,\mu\rangle) = \langle Af,\mu\rangle \tag{6.8}$$

Then using this together with the collapsing sum

$$
\begin{aligned}
&a_1\dots a_n + (a_1\dots a_{n-1}b_n - a_1\dots a_n) + (a_1\dots a_{n-2}b_{n-1}b_n - a_1\dots a_{n-1}b_n) \\
&+ (b_1\dots b_n - a_1b_2\dots b_n) - b_1\dots b_n = 0 \\
&a_1\dots a_n - b_1\dots b_n = \sum_k(a_1\dots a_kb_{k+1}\dots b_n - a_1\dots a_{k+1}b_{k+1}\dots b_n).
\end{aligned}
$$

or rewriting

$$\prod_{j=1}^n\langle f,\mu P_N\rangle = \prod_{j=1}^n[(\langle f,\mu\rangle + (\langle f,\mu P_N\rangle - \langle f,\mu\rangle))] \tag{6.9}$$

we obtain

$$\begin{aligned} NE_\mu[F(X_1^N) - F(\mu)] &= \sum_{1\le i<j\le n} \left(\langle f_i f_j, \mu P_N\rangle - \langle f_i, \mu P_N\rangle \langle f_j, \mu P_N\rangle\right) \prod_{\ell:\ell\neq i,j} \langle f_\ell, \mu P_N\rangle \\ &\quad + \sum_{i=1}^{n} \langle Af_i, \mu\rangle \prod_{j:j<i} \langle f_j, \mu\rangle \prod_{j:j>i} \langle f_j, \mu P_N\rangle + O(N^{-1}) \\ &= GF(\mu) + o(1) \end{aligned}$$

uniformly in μ.

The completes the verification of condition (6.7). ■

6.4 The Infinitely Many Alleles Model

This is a special case of the Fleming-Viot process which has played a crucial role in modern population biology. It has type space $E = [0, 1]$ and *type-independent* mutation operator with mutation source $\nu_0 \in \mathcal{P}([0, 1])$

$$\begin{aligned} Af(x) &= \theta(\int p(x, dy)f(y) - f(x)) \\ &= \theta(\int f(y)\nu_0(dy) - f(x)). \end{aligned}$$

Since A is a bounded operator we can take indicator functions of intervals in $D(A)$. If we have a partition $[0, 1] = \cup_{j=1}^{K} B_j$ where the B_j are intervals, consider the set $D(G)$ of functions

$$F(\mu) = \langle f_1, \mu\rangle \dots \langle f_n, \mu\rangle \tag{6.10}$$

with $n \ge 1$ and where the functions $f_1, \dots, f_n$ are finite linear combinations of indicator functions of the intervals $\{B_j\}$. Then the function $GF(\mu)$ can be written in the same form and we can prove that the Δ_{K-1}-valued process $\{p_t(B_1), \dots, p_t(B_K)\}$ is a version of the $K - allele$ process with generator

$$\begin{aligned} &G^K f(p) \\ &= \frac{1}{2} \sum_{i,j=1}^{K-1} p_i(\delta_{ij} - p_j) \frac{\partial^2 f(p)}{\partial p_i \partial p_j} + \theta \sum_{i=1}^{K-1} (\nu_0(B_i) - p_i) \frac{\partial f(p)}{\partial p_i}. \end{aligned} \tag{6.11}$$

We will next give an explicit construction of this process that allows us to derive a number of interesting properties of this important model.