Dirichlet and Poisson-Dirichlet approximation of genetic drift models

Han Liang Gan

Northwestern University

March, 2018

Joint work with Adrian Röllin (Singapore) and Nathan Ross (Melbourne).

The Dirichlet distribution

We will define the Dirichlet distribution on the K-simplex (Δ_K) as a $K-1$ dimensional distribution with parameters $(a_1, a_2, \ldots, a_K)$ and density

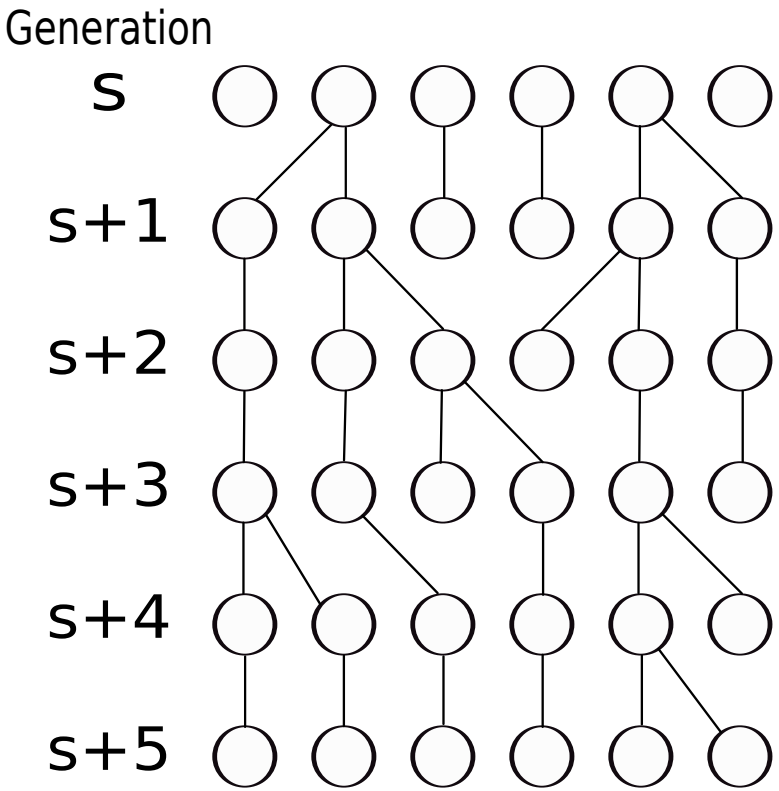
$$\frac{\Gamma\left(\sum_{i=1}^{K} a_i\right)}{\prod_{i=1}^{K} \Gamma(a_i)} \prod_{i=1}^{K} x_i^{a_i - 1},$$

where $x_i \in [0,1]$, $\sum_{i=1}^{K} x_i = 1$.

Note that in all of the following we will set $s = \sum_{i=1}^{K} a_i$.

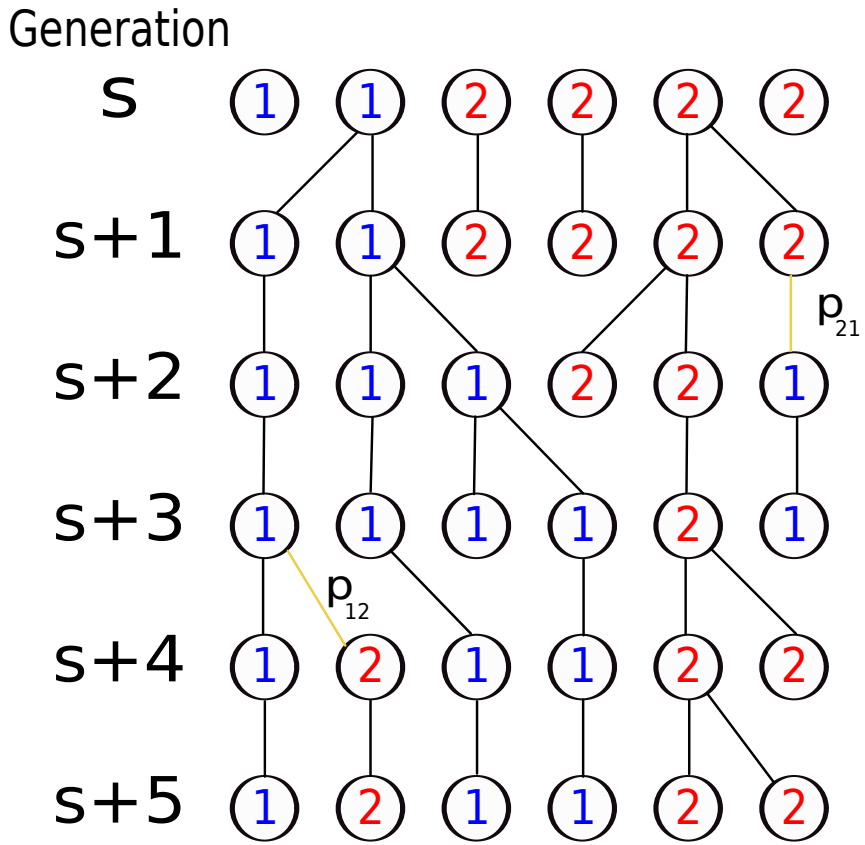
The Wright-Fisher model

- The model used to model gene frequencies as the change from generation to generation.
- We assume fixed population size N in each non-overlapping generation.
- Children select parents uniformly at random.



Mutation

- After selecting a parent, each child takes the same type as the parent.
- There is then a chance that it mutates to a different type.
- p_{ij} is the probability of mutating from type i to j.



Limit results for the stationary distributions

As $t \to \infty$, let $\mathbf{W} = (W_1, W_2, \ldots, W_K)$ denote the proportions of the K types in the stationary distribution for the Wright-Fisher model.

It is known that assuming a parent independent mutation (PIM) structure, that is $p_{ij} = p_j$, then the scaling limit of $\mathbf{W}$ is Dirichlet distribution.

What can we say for some arbitrary mutation structure p_{ij}?

Wright-Fisher model with general mutation structure

Theorem

For any three times partially differentiable function $h : \Delta_K \to \mathbb{R}$,

$$|\mathbb{E}h(\boldsymbol{W}) - \mathbb{E}h(\boldsymbol{Z})| \leq \frac{|h|_1}{s} A_1 + \frac{|h|_2}{2(s+1)} A_2 + \frac{|h|_3}{18(s+2)} A_3,$$

where $\boldsymbol{Z} \sim Dir(\boldsymbol{a})$ *and*

$$A_1 = 2N(K+1)\tau, \quad A_2 = NK^2\mu^2 + 2K\mu, \quad A_3 = 8NK^3\mu^3 + \frac{16\sqrt{2}K^3}{N^{1/2}},$$

with

$$\tau = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} |\, p_{ij} - \frac{a_j}{2N}|, \qquad \mu = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} p_{ij}.$$

Theorem (cont.)

Moreover, there is a constants $C = C(\boldsymbol{a})$ *and* $\theta = \theta(\boldsymbol{a}) > 0$ *such that*

$$\sup_{A \in C_{K-1}} |\mathbb{P}[\boldsymbol{W} \in A] - \mathbb{P}[\boldsymbol{Z} \in A]| \leq C \left(A_1 + A_2 + A_3\right)^{\theta/(3+\theta)},$$

where C_{K-1} *is the family of convex sets on* $\mathbb{R}^{K-1}$.

Take home messages

If we assume PIM, the smooth test function bound is of order $N^{-1/2}$ and the convex set metric is of order $N^{-1/8}$ as long as all the $a_i \geq 1$.

If we don't have PIM structure, these bounds give us explicit error bounds. In particular, if $p_{ij} = a_j/2N + \epsilon_{ij}$ and $\epsilon = o(N^{-1})$ then the bound goes to zero.

PIM Cannings model

In the *Cannings model*, the distribution of the selection is said to be *exchangeable*. For the PIM Cannings model we have a similar result to the Wright-Fisher model.

The bound for smooth test functions is of order $N^{-1/2}$ and the bound for the convex set metric is $N^{-1/8}$ when the $a_i \geq 1$.

Stein's method for the Dirichlet distribution

Lemma

Let $a_1, \ldots, a_K$ be positive numbers and $s = \sum_{i=1}^K a_i$. The random vector $\boldsymbol{W} \in \Delta_K$ has distribution $\mathrm{Dir}(a_1, \ldots, a_K)$ if and only if for all $f : \Delta_K \to \mathbb{R}$ bounded with two bounded derivatives

$$\mathbb{E}\mathcal{A}f(\boldsymbol{W}) := \mathbb{E}\left[\sum_{i,j=1}^{K-1} W_i(\delta_{ij} - W_j)f_{ij}(\boldsymbol{W}) + \sum_{i=1}^{K-1}(a_i - sW_i)f_i(\boldsymbol{W})\right] = 0.$$

A different mutation regime

Does it make sense that there is a fixed number of types that the genes can take?

When we think of mutations, we typically think of this meaning that something *new* has appeared.

What if mutations lead only to new types?

Poisson-Dirichlet distribution

Arguably the simplest way to think about what the Poisson-Dirichlet distribution is via what is known as the Chinese restaurant process.

Another alternative is as the scaling limit of the stationary distribution for the Wright-Fisher model with mutation, but where mutation always goes to a new type. This sort of model is often called the infinite-alleles model.

Our goal

We want to see if we can prove results for Poisson-Dirichlet approximation that are analogous to the results we have for Dirichlet approximation.

It turns out that this is surprisingly complicated.

Our space of random measures

For technical reasons, we need to actually consider our type space to be $U[0,1]$, and hence mutations change to a uniformly chosen at random value between 0 and 1. (This is also known as a Fleming-Viot process.)

What we have is a purely atomic random measure on $U[0,1]$. And it turns out convergence of measures in this space is tricky.

What can we show

Weak convergence (including explicit bounds on rates of convergence) of the stationary measure of a finite N infinite alleles model to the stationary distribution of the Fleming-Viot diffusion.

What haven't we shown

In a sense, what we have shown is weak convergence of point (diffusion) processes.

This therefore includes both location and size information.

Our original goal was just for Poisson-Dirichlet approximation, did we achieve this goal?