On Selective Sweeps with Recombination

Adi Krishnamoorthy

June 2025

1 Introduction

We define a selective sweep [1] as the process in which a new allele in an individual resulting from a beneficial mutation in the individual's genome can spread through the entire population of individuals over a period of time. In this paper, we propose an approximation for a selective sweep based on Markov Chains and Differential Equations.

In the classical Moran model [2] for population dynamics, we have a population (or collection) of individuals with a fixed size of N. At a single site on a chromosome, each individual can have two alleles (or different copies of a gene). Now, each individual in the population can live for an amount of time that is exponentially distributed with parameter λ, after which it dies. At this point, another individual is born, with its parent chosen randomly from the population. The newly birthed individual inherits the same copy of the allele as its parent individual.

In this paper, we examine a similar model to the classical Moran model where we model two different loci on a chromosome - one where a beneficial mutation occurs and another, neutral allele situated at a certain distance away from the first locus. We assume that at the site where the beneficial mutation occurs, which we call the selected site, the two possible alleles are denoted as type-0 and type-1, where 0 represents the beneficial (mutated) allele and 1 represents the ancestral (original) allele. Furthermore, at the neutral site, we also label the gene there as type-0 or type-1, where type-0 represents that the gene was descended from the individual that acquired the beneficial mutation and type-1 represents that it was descended from one of the other chromosomes at the start of the sweep.

We define selection to be a notion of how advantageous the beneficial allele is over the ancestral allele when new individuals are being born. For all N, we denote s_N as the selection probability, where $1/N \ll s_N \ll 1$. We choose s_N such that for some $0 < \alpha < 1$, as $N \to \infty$, $s_n N^\alpha \to s$. Then, we let the relative fitnesses of the 0 and 1 alleles be 1 and $1 - s_N$. We then follow the same, Moran model as before, except we reject the replacement of a gene with a type-0 allele at the selected site by a gene with the type-1 allele at the selected site with probability $1 - s_N$.

In this model, we also take the notion of recombination into account. Recombination is a biological process where chromosome pieces in parent individuals are broken and then rearranged so new allele combinations can be produced, and the birthed individual inherits those alleles. In this model, each individual in the population is really a chromosome, so recombination here would mean that a newly birthed individual can inherit the gene at the selected site from one parent and the gene at the neutral site from another parent. Because of this, we examine the alleles at the selected and neutral site separately, and hence we model it with a multi-dimensional Markov chain. Now, the

recombination probability is proportional to the distance between the seleccted and neutral sites. For all N, we let this probability be equal to r_N, where $1/N \ll r_N$. Now, we choose r_N such that for some $\alpha > 0$, $r_N N^\alpha \to r$ as $N \to \infty$, where r is a positive real number. We then examine this model under the weak genetic draft [3] regime, a regime in which r and s are related in that $0 < r < s$.

We divide this paper into two sections. In the first section, we give a more in-depth description of the model we use for the selective sweep with recombination under the weak genetic draft regime. We use an ordinary differential equation to approximate a continuous-time Markov chain modeling the middle of the selective sweep. In the second section, we describe the beginning and end portions of the sweep.

2 Sweep Model

2.1 Results from Darling-Norris (2007)

One of the core principles of *Differential Equation Approximations for Markov Chains [4]* is to understand when the paths of a Markov chain $(X_t)_{t\geq 0}$ will be close to the solution of a differential equation with high probability. To show the convergence of this Markov chain, we define a drift vector, $b(x(t))$, and set it equal to the product of the average jump size and the expected rate of the jumps. We then set the limit of the Markov chain to be equal to the solution of the differential equation $x'(t) = b(x(t))$.

To prove this concept, we use Theorem 4.1 from [4]. If $(\boldsymbol{X_t})_{t\geq 0} = (X_t^1, ..., X_t^d)$ is a d-dimensional continuous time Markov chain and $x = (x^1, ..., x^d) : S \to \mathbb{R}^d$ is a set of coordinate functions, then for all $\xi \in S$ (where S is the state space), we define the drift vector as

$$\beta(\xi) = \sum_{\xi' \neq \xi} (x(\xi') - x(\xi)) q(\xi, \xi').$$

Suppose $x_0 \in U \subseteq \mathbb{R}^d$ and $b : U \to \mathbb{R}^d$ is a Lipschitz vector field, and let $T_1 = \inf\{t \geq 0 : \beta(X_t) = \infty\}$. Define $(M_t)_{t\in[0,T_1]}$ such that

$$\boldsymbol{X_t} = \boldsymbol{X_0} + M_t + \int_0^t \beta(X_s) ds, \quad 0 \leq t \leq T_1$$

and define x_t such that

$$x_t = x_0 + \int_0^t b(x_s) ds, \quad 0 \leq t.$$

Now, we let $t_0 < \zeta$, $\epsilon > 0$, $A > 0$ and $\delta = \epsilon e^{-Kt_0}/3$ (where K is the Lipschitz constant for b on U). For all $\xi \in S$, we let $\alpha(\xi) = \sum_{\xi' \neq \xi} |x(\xi') - x(\xi)|^2 q(\xi, \xi')$. Then, theorem 4.1 of [4] states that:

$$P\left(\sup_{t\leq t_0} |X_t - x_t| > \epsilon\right) \leq \frac{4At_0}{\delta^2} + P\left(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c\right),$$

where $\Omega_0, \Omega_1, \Omega_2$ are events such that $\Omega_0 = \{|\boldsymbol{X_0} - x_0| \leq \delta\}$, $\Omega_1 = \{\int_0^{T\wedge t_0} |\beta(X_t) - b(\boldsymbol{x}(X_t))| dt \leq \delta\}$, and $\Omega_2 = \{\int_0^{t\wedge T_0} \alpha(X_t) dt \leq At_0\}$.

In the rest of this section, we derive a differential equation approximation for the selective sweep using the techniques above. We define a continuous-time Markov Chain showing the movement of the evolutionary process. We then use the previous technique to approximate these Markov Chains by differential equations.

2.2 One locus

We first create an approximation for the number of individuals with the type-0 (beneficial) allele at the selected site during the course of the sweep. Suppose that $(\tilde{X}_t)_{t\geq 0}$ is a continuous-time Markov chain representing the number of lineages with the type-0 allele at the selected site. Let $S = \{0, 1, ..., N\}$ be the state space. If the chain is in state i, it means that i lineages have the type-0 allele. We let $\tilde{q}_{i,j}$ represent the transition rate from state i to state j.

When the Markov Chain is in state i it could jump to either state $i+1$ or state $i-1$. The chain will jump from i to $i+1$ if a type-1 individual dies, the new individual has a type-0 allele, and the change is accepted. The chain will jump from i to $i-1$ if a type-0 individual dies, the new individual has a type-1 allele, and the change is accepted. Note that since the relative fitness of the type 0 allele to the type 1 allele is 1 to $1-s_N$, the probability that a change from a type-1 to a type-0 allele is accepted is 1 and the probability that a change from a type-0 to a type-1 allele is accepted is $1-s_N$.

The number of alleles that are of type 1 when the Markov chain is in state i equals $N-i$ and the probability that the parent of a new allele will be of type 0 is i/N. The probability that the change is accepted will be 1. This gives us:

$$\tilde{q}(i, i+1) = (N-i)\frac{i}{N} \cdot 1 = \frac{i(N-i)}{N}.$$

Similarly, the number of alleles that are of type 0 equals i and the probability that the parent of a new allele will be of type 0 is $(N-i)/N$. The probability that the change is accepted will be $1-s$. So,

$$\tilde{q}(i, i-1) = i \cdot \frac{N-i}{N} \cdot (1-s) = \frac{i(N-i)}{N}(1-s).$$

Now, we rescale $\tilde{X}_t$ to a scale over the interval $[0, 1]$. Define $X_t = \tilde{X}_{N^\alpha t}/N$. For all $N \in \mathbb{N}$, we let $s_N = sN^{-\alpha}$ be the rescaled selection probability. Then, $S = \{0, 1/N, 2/N, ..., (N-1)/N, 1\}$ will be the new state space. The rescaled transition rates will be:

$$q(i/N, (i+1)/N) = N^\alpha \frac{i(N-i)}{N},$$

$$q(i/N, (i-1)/N) = N^\alpha \frac{i(N-i)}{N}(1-s_N).$$

Now, we use the concept of the drift vector, which we previously defined as a differential equation where the rate of change at a certain time is equal to the average jump rate of this Markov chain. We then use a fluid limit [4,5] to show convergence of the Markov chain to this differential equation.

Theorem 2.1. *Suppose that $\epsilon > 0$ and $X_0 = \lfloor \epsilon N \rfloor / N$. Define the function x_t, where $x_t' = sx_t(1-x_t)$ and $x_0 = \epsilon$ as $N \to \infty$. Then, for all $t_0 > 0$ and $\eta > 0$, for sufficiently large N we have*

$$P\left(\sup_{t \leq t_0} |X_t - x_t| > \epsilon\right) \leq \eta.$$

Proof. Given N, the drift vector for the Markov chain $(X_t)_{t\geq 0}$ is

$$\begin{aligned}\beta_N\left(\frac{i}{N}\right) &= \frac{1}{N}\cdot q\left(\frac{i}{N},\frac{i+1}{N}\right)+\frac{(-1)}{N}\cdot q\left(\frac{i}{N},\frac{i-1}{N}\right)\\ &= N^\alpha\frac{i(N-i)}{N^2}-N^\alpha\frac{i(N-i)}{N^2}(1-s_N)\\ &= N^\alpha\frac{i(N-i)}{N^2}(1-1+s_N)\\ &= s_N N^\alpha\frac{i}{N}\left(\frac{N-i}{N}\right).\end{aligned}$$

Note that we have $s_N N^\alpha \to s$ as $N\to\infty$. So, we let the drift vector be $b(x_t)=sx_t(1-x_t)$. Let $x_0=\epsilon$ for some $\epsilon>0$ and $x_t'=b(x_t)$.

To prove convergence, we use the previous theorem described in Section 2.1. We know that $(X_t)_{t\geq 0}$ is the Markov chain. Then if $\delta=\epsilon e^{-Kt_0}/3$, the quantity $4At_0/\delta^2$ will go to 0 if $A\to 0$. So for all A, we have

$$P(\sup_{t\leq t_0}|X_t-x_t|>\epsilon)\leq 4At_0/\delta^2+P(\Omega_0^c\cup\Omega_1^c\cup\Omega_2^c).$$

Recall that $x_0=\epsilon$ and $X_0=\lfloor\epsilon N\rfloor/N$. Then, $|X_0-x_0|\leq 1/N<\delta$ as $N\to\infty$. So for sufficiently large N, $P(|X_0-x_0|<\delta)=1$, and so $P(\Omega_0)=1$. Thus $P(\Omega_0^c)=0$.

Now, for all i, $|\beta_N(i/N)-b(i/N)|=|s_N N^\alpha i(N-i)/N^2-si(N-i)/N^2|\to 0$ as $N\to\infty$ (since $\lim_{n\to\infty}s_N N^\alpha=s$). So for sufficiently large N,

$$\int_0^{t_0}|\beta_N(X_t)-b_N(X_t)|dt\leq\delta$$

and so $P(\Omega_1)=P(\int_0^{t_0}|\beta_N(i/N)-b(i/N)|dt\leq\delta)=1$ for sufficiently large N. Therefore, $P(\Omega_1^c)=0$ for sufficiently large N. Now, suppose $A>0$. Then,

$$\alpha\left(\frac{i}{N}\right)=\sum_{i'\neq i}\left|\frac{i}{N}'-\frac{i}{N}\right|^2 q\left(\frac{i}{N},\frac{i'}{N}\right)=(2-s_N)\frac{1}{N^2}N^\alpha\frac{i(N-i)}{N}$$

so

$$\int_0^{t_0}\alpha(X_t)dt\leq(2-s_N)\int_0^{t_0}\frac{1}{N^2}N^\alpha\frac{X_t(N-X_t)}{N}dt\to 0\text{ as }N\to\infty.$$

This means that for sufficiently large N, $P(\Omega_2)=P(\int_0^{t_0}\alpha(X_t)dt\leq At_0)=1$. Therefore $P(\Omega_2^c)=0$, and hence for sufficiently large N, we have $P(\Omega_0^c\cup\Omega_1^c\cup\Omega_2^c)=0$.

This means that for any fixed A and large N, we have

$$P\left(\sup_{t\leq t_0}|X_t-x_t|>\epsilon\right)\leq 4At_0/\delta^2 \qquad (\star)$$

and since A can be arbitrarily small,

$$P\left(\sup_{t\leq t_0}|X_t-x_t|>\epsilon\right)\leq\eta$$

for all $\eta>0$. □

We know that the process starts at ϵ. Then, the initial value of x_t at time 0 is ϵ, and so the logistic equation modeling this process will be:

$$x_t = \frac{1}{1 + \frac{1-\epsilon}{\epsilon} e^{-st}}. \tag{1}$$

which is the solution to the differential equation in Theorem 2.1.

2.3 Two loci

We now extend this model to the two-locus scenario - one being the selected site and one being the neutral site. In this scenario, each individual (in the population of size N) can be in one of 4 states:

$(0,0)$: the individual has the type-0 allele at the selected site and the type-0 allele at the neutral site
$(0,1)$: the individual has the type-0 at the selected site and the type-1 allele at the neutral site
$(1,0)$: the individual has the type-1 allele at the selected site and the type-0 allele at the neutral site
$(1,1)$: the individual has the type-1 at the selected site and the type-1 allele at the neutral site.

We let $\tilde{W}_t, \tilde{X}_t, \tilde{Y}_t$, and $\tilde{Z}_t$ be the number of individuals in the population in the $(0,0), (0,1), (1,0)$, and $(1,1)$ states respectively at time t. Clearly, $\tilde{W}_t + \tilde{X}_t + \tilde{Y}_t + \tilde{Z}_t = N$ as there are a total of N individuals and each can be in exactly one of these states. We extend the previous Markov Chain to a three dimensional one, where the state space is $\{0, ..., N\}^4$ and if the chain is state (i, j, k, l) at time t, then $\tilde{W}_t = i$, $\tilde{X}_t = j$, $\tilde{Y}_t = k$, and $\tilde{Z}_t = l$.

Now, suppose that $0 < \epsilon < 1$ is arbitrary. In our model, we assume that we start with one $(0,0)$ individual and that at first, the number of individuals with the type-0 allele in the selected site grows approximately like a supercritical branching process with rate s. Then after a short time t, the number of individuals with the type-0 allele at the selected site is approximately e^{st}. So initially, $\tilde{W}_t + \tilde{X}_t \propto e^{st} \propto N$. On the other hand, the number of $(0,0)$ individuals grows approximately like a supercritical branching process with rate $s - r$. Therefore, after a short time t, the number of individuals with the type-0 allele at the neutral site is approximately $e^{(s-r)t}$ So, we let $\beta = 1 - r/s$ and $\tilde{W}_t + \tilde{Y}_t \propto e^{(s-r)t} \propto N^\beta$. Furthermore, we assume that $r/s < (1-\alpha)/2$.

Then, we define the initial conditions to be after the short time frame described above. Let $\tilde{W}_0 = N^\beta$, $\tilde{X}_0 = \lfloor \epsilon N \rfloor$, $\tilde{Y}_0 = 0$, and $\tilde{Z}_0 = 1 - \lfloor \epsilon N \rfloor - N^\beta$.

At any given time, there are 12 possible transitions that can be made. We will formulate the first one as follows:

$$\tilde{q}((i,j,k,l),(i+1,j-1,k,l)) = (1 - r_N)(j)\frac{(i)}{N} + r_N(j)\frac{(i+j)}{N}\frac{(i+k)}{N}$$

Here, the first term arises from the scenario where no recombination occurs. Here, one of the $(0,1)$ individuals dies and is replaced by a $(0,0)$ individual. The second term arises from the scenario where recombination occurs. Here, one of the $(0,1)$ individuals dies and is replaced by an individual

with a 0 allele at both sites. Simplifying this equation, we get

$$\tilde{q}((i,j,k,l),(i+1,j-1,k,l)) = j\left[\frac{i}{N}(1-r_N) + r_N\left(\frac{i+j}{N}\right)\left(\frac{i+k}{N}\right)\right]$$

which is this transition probability. We use the same method to calculate the other transition probabilities, as follows:

$$\tilde{q}((i,j,k,l),(i+1,j,k-1,l)) = k\left[\frac{i}{N}(1-r_N) + r_N\left(\frac{i+j}{N}\right)\left(\frac{i+k}{N}\right)\right]$$

$$\tilde{q}((i,j,k,l),(i+1,j,k,l-1)) = l\left[\frac{i}{N}(1-r_N) + r_N\left(\frac{i+j}{N}\right)\left(\frac{i+k}{N}\right)\right]$$

$$\tilde{q}((i,j,k,l),(i-1,j+1,k,l)) = i\left[\frac{j}{N}(1-r_N) + r_N\left(\frac{i+j}{N}\right)\left(\frac{j+l}{N}\right)\right]$$

$$\tilde{q}((i,j,k,l),(i,j+1,k-1,l)) = k\left[\frac{j}{N}(1-r_N) + r_N\left(\frac{i+j}{N}\right)\left(\frac{j+l}{N}\right)\right]$$

$$\tilde{q}((i,j,k,l),(i,j+1,k,l-1)) = l\left[\frac{j}{N}(1-r_N) + r_N\left(\frac{i+j}{N}\right)\left(\frac{j+l}{N}\right)\right]$$

$$\tilde{q}((i,j,k,l),(i-1,j,k+1,l)) = i\left[\frac{k}{N}(1-r_N) + r_N\left(\frac{i+k}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$\tilde{q}((i,j,k,l),(i,j-1,k+1,l)) = j\left[\frac{k}{N}(1-r_N) + r_N\left(\frac{i+k}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$\tilde{q}((i,j,k,l),(i,j,k+1,l-1)) = l\left[\frac{k}{N}(1-r_N) + r_N\left(\frac{i+k}{N}\right)\left(\frac{k+l}{N}\right)\right]$$

$$\tilde{q}((i,j,k,l),(i-1,j,k,l+1)) = i\left[\frac{l}{N}(1-r_N) + r_N\left(\frac{j+l}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$\tilde{q}((i,j,k,l),(i,j-1,k,l+1)) = j\left[\frac{l}{N}(1-r_N) + r_N\left(\frac{j+l}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$\tilde{q}((i,j,k,l),(i,j,k-1,l+1)) = k\left[\frac{l}{N}(1-r_N) + r_N\left(\frac{j+l}{N}\right)\left(\frac{k+l}{N}\right)\right]$$

We then rescale these equations. Suppose that $\alpha \in (0,1)$ and β are as defined previously.

We let W_t, X_t, Y_t, and Z_t be the rescaled versions of $\tilde{W}_t, \tilde{X}_t, \tilde{Y}_t$, and $\tilde{Z}_t$. Then, $W_t = \tilde{W}_{N^\alpha t}/N^\beta, X_t =$

$\tilde{X}_{N^\alpha t}/N, Y_t = \tilde{Y}_{N^\alpha t}/N^\beta$, and $Z_t = \tilde{Z}_{N^\alpha t}/N$. So the rescaled transition rates will be:

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i+1}{N^\beta},\frac{j-1}{N},\frac{k}{N^\beta},\frac{l}{N}\right)\right) = N^\alpha j\left[\frac{i}{N}(1-r_N)+r_N\left(\frac{i+j}{N}\right)\left(\frac{i+k}{N}\right)\right]$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i+1}{N^\beta},\frac{j}{N},\frac{k-1}{N^\beta},\frac{l}{N}\right)\right) = N^\alpha k\left[\frac{i}{N}(1-r_N)+r_N\left(\frac{i+j}{N}\right)\left(\frac{i+k}{N}\right)\right]$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i+1}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l-1}{N}\right)\right) = N^\alpha l\left[\frac{i}{N}(1-r_N)+r_N\left(\frac{i+j}{N}\right)\left(\frac{i+k}{N}\right)\right]$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i-1}{N^\beta},\frac{j+1}{N},\frac{k}{N^\beta},\frac{l}{N}\right)\right) = N^\alpha i\left[\frac{j}{N}(1-r_N)+r_N\left(\frac{i+j}{N}\right)\left(\frac{j+l}{N}\right)\right]$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i}{N^\beta},\frac{j+1}{N},\frac{k-1}{N^\beta},\frac{l}{N}\right)\right) = N^\alpha k\left[\frac{j}{N}(1-r_N)+r_N\left(\frac{i+j}{N}\right)\left(\frac{j+l}{N}\right)\right]$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i}{N^\beta},\frac{j+1}{N},\frac{k}{N^\beta},\frac{l-1}{N}\right)\right) = N^\alpha l\left[\frac{j}{N}(1-r_N)+r_N\left(\frac{i+j}{N}\right)\left(\frac{j+l}{N}\right)\right]$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i-1}{N^\beta},\frac{j}{N},\frac{k+1}{N^\beta},\frac{l}{N}\right)\right) = N^\alpha i\left[\frac{k}{N}(1-r_N)+r_N\left(\frac{i+k}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i}{N^\beta},\frac{j-1}{N},\frac{k+1}{N^\beta},\frac{l}{N}\right)\right) = N^\alpha j\left[\frac{k}{N}(1-r_N)+r_N\left(\frac{i+k}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k+1}{N^\beta},\frac{l-1}{N}\right)\right) = N^\alpha l\left[\frac{k}{N}(1-r_N)+r_N\left(\frac{i+k}{N}\right)\left(\frac{k+l}{N}\right)\right]$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i-1}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l+1}{N}\right)\right) = N^\alpha i\left[\frac{l}{N}(1-r_N)+r_N\left(\frac{j+l}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i}{N^\beta},\frac{j-1}{N},\frac{k}{N^\beta},\frac{l+1}{N}\right)\right) = N^\alpha j\left[\frac{l}{N}(1-r_N)+r_N\left(\frac{j+l}{N}\right)\left(\frac{k+l}{N}\right)\right](1-s_N)$$

$$q\left(\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right),\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k-1}{N^\beta},\frac{l+1}{N}\right)\right)=N^\alpha k\left[\frac{l}{N}(1-r_N)+r_N\left(\frac{j+l}{N}\right)\left(\frac{k+l}{N}\right)\right]$$

We then use the theorem in section 2.1 to show convergence to a set of differential equations.

Theorem 2.2. *Suppose $\theta,\epsilon>0$. Given the Markov chains $(W_t)_{t\geq 0},(X_t)_{t\geq 0},(Y_t)_{t\geq 0},(Z_t)_{t\geq 0}$, suppose $W_0=\lfloor\theta N^\beta\rfloor/N^\beta$, $X_0=\lfloor\epsilon N\rfloor/N$, $Y_0=0$, and $Z_0=((N-\lfloor\epsilon N\rfloor)/N)-(\lfloor\theta N^\beta\rfloor/N^\beta)$. Then, define:*

$$\begin{aligned}
w_t' &= sw(y+z)+r(xy-wz)\\
x_t' &= sx(y+z)+r(wz-xy)\\
y_t' &= -s(1-r)y(w+x)-sr(w+y)(y+z)(w+x)+r(w+y)(y+z)-ry\\
z_t' &= -s(1-r)z(w+x)-sr(x+z)(y+z)(w+x)+r(x+z)(y+z)-rz,
\end{aligned}$$

with $w_0=\theta,x_0=\epsilon,y_0=0$, and $z_0=1-\epsilon$. Then,

$$P\left(\sup_{t\leq t_0}|W_t-w_t|>\epsilon\right)\leq\eta$$

$$P\left(\sup_{t\leq t_0}|X_t-x_t|>\epsilon\right)\leq\eta$$

$$P\left(\sup_{t\leq t_0}|Y_t-y_t|>\epsilon\right)\leq\eta$$

$$P\left(\sup_{t\leq t_0}|Z_t-z_t|>\epsilon\right)\leq\eta$$

for all $\eta>0$.

Proof. We can then write the Markov chains for W_t,X_t,Y_t, and Z_t as follows:

$$\begin{aligned}
&\beta_{W,N}\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right)\\
&=\frac{1}{N^\beta}(q((i,j,k,l),(i+1/N,j-1/N,k,l))+q((i,j,k,l),(i+1/N,j,k-1/N,l))\\
&\quad+q((i,j,k,l),(i+1/N,j,k,l-1/N))-q((i,j,k,l),(i-1/N,j+1/N,k,l))\\
&\quad-q((i,j,k,l),(i-1/N,j,k+1/N,l))-q((i,j,k,l),(i-1/N,j,k,l-1/N)))\\
&=\frac{N^\alpha}{N^\beta}\left(\frac{s_N}{N}i(k+l)+\frac{r_N}{N}(jk-il)\right)
\end{aligned}$$

Now, for large N, $N^\alpha s_N\approx s$ and $N^\alpha r_N\approx r$. Also, since i and k are of order N^β, for large N, $i/N\approx 0$ and $k/N\approx 0$. So, the above equation can be approximated as follows:

$$\frac{N^\alpha}{N^\beta}\left(\frac{s_N}{N}i(k+l)+\frac{r_N}{N}(jk-il)\right)\approx s\left(\frac{i}{N^\beta}\cdot\frac{l}{N}\right)-r\left(\frac{j}{N}\cdot\frac{k}{N^\beta}-\frac{i}{N^\beta}\cdot\frac{l}{N}.\right)$$

Similarly,

$$\begin{aligned}
\beta_{X,N}\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right) &= N^\alpha\left(\frac{s_N}{N^2}j(k+l)+\frac{r_N}{N^2}(il-jk)\right)\approx s\left(\frac{j}{N}\cdot\frac{l}{N}\right)\\
\beta_{Y,N}\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right) &= \frac{N^\alpha}{N^\beta}(-\frac{s_N}{N}(1-r_N)k(i+j)-\frac{s_N}{N^2}r_N(i+k)(k+l)(i+j)\\
&\quad+\frac{r_N}{N}(i+k)(k+l)-r_Nk)\\
&\approx -s\left(\frac{k}{N^\beta}\cdot\frac{j}{N}+r\left(\frac{i+k}{N^\beta}\cdot\frac{l}{N}\right)-r\frac{k}{N^\beta}\right)\\
\beta_{Z,N}\left(\frac{i}{N^\beta},\frac{j}{N},\frac{k}{N^\beta},\frac{l}{N}\right) &= N^\alpha(-\frac{s_N}{N^2}(1-r_N)l(i+j)-\frac{s_N}{N^3}r_N(j+l)(k+l)(i+j)\\
&\quad+\frac{r_N}{N^2}(j+l)(k+l)-\frac{r_N}{N}l)\\
&\approx -s\left(\frac{l}{N}\cdot\frac{j}{N}\right)
\end{aligned}$$

So, we define the drift functions as follows:

$$\begin{aligned}
b_w(w_t,x_t,y_t,z_t) &= swz+r(xy-wz)\\
b_x(w_t,x_t,y_t,z_t) &= sxz\\
b_y(w_t,x_t,y_t,z_t) &= -syx+r(w+y)z-ry\\
b_z(w_t,x_t,y_t,z_t) &= -szx
\end{aligned}$$

We define w_t,x_t,y_t,z_t as functions such that $w_t' = b_w(w_t,x_t,y_t,z_t), x_t' = b_x(w_t,x_t,y_t,z_t), y_t' = b_y(w_t,x_t,y_t,z_t)$, and $z_t' = b_z(w_t,x_t,y_t,z_t)$ (where $w_0=\theta>0$, $x_0=\epsilon, y_0=0$ and $z_0=1-\epsilon$). To show convergence of $(W_t),(X_t),(Y_t),(Z_t)$ to w,x,y,z, we use the same theorem as before. We know that $(W_t,X_t,Y_t,Z_t)_{t\geq 0}$ is the Markov Chain being measured. Then, as $\epsilon\to 0$, for $\delta=\epsilon e^{-Kt_0}/3$, and for sufficiently small A, we have

$$\begin{aligned}
P\left(\sup_{t\leq t_0}|W_t-w_t|>\epsilon\right) &\leq 4At_0/\delta^2+P(\Omega_0^c\cup\Omega_1^c\cup\Omega_2^c),\\
P\left(\sup_{t\leq t_0}|X_t-x_t|>\epsilon\right) &\leq 4At_0/\delta^2+P(\Omega_0^c\cup\Omega_1^c\cup\Omega_2^c),\\
P\left(\sup_{t\leq t_0}|Y_t-y_t|>\epsilon\right) &\leq 4At_0/\delta^2+P(\Omega_0^c\cup\Omega_1^c\cup\Omega_2^c),\\
P\left(\sup_{t\leq t_0}|Z_t-z_t|>\epsilon\right) &\leq 4At_0/\delta^2+P(\Omega_0^c\cup\Omega_1^c\cup\Omega_2^c).
\end{aligned}$$

Clearly, if the process starts where $w_0=\epsilon$ and $z_0=1-\epsilon$, with $W_0=\lfloor\theta N^\beta\rfloor/N^\beta$, $X_0=\lfloor\epsilon N\rfloor/N$, $Y_0=0$, and $Z_0=((N-\lfloor\epsilon N\rfloor)/N)-(\lfloor\theta N^\beta\rfloor/N^\beta)$, then $P(\Omega_0)=1<\delta$, so $P(\Omega_0^c)=0$.

We now bound $P(\Omega_1)$. For sufficiently large N, we check each of the four chains. We see that:

$$\int_0^{t_0}\left|\beta_{W,N}(\tilde{W}_t,\tilde{X}_t,\tilde{Y}_t,\tilde{Z}_t)-b_w(W_t,X_t,Y_t,Z_t)\right|dt$$

$$= \int_0^{t_0} \left| \frac{N^\alpha}{N^\beta} \left(\frac{s_N}{N} \tilde{W}_t (\tilde{Y}_t + \tilde{Z}_t) + \frac{r_N}{N} \left(\tilde{X}_t \tilde{Y}_t - \tilde{W}_t \tilde{Z}_t \right) \right) - s W_t Z_t - r X_t Y_t + r W_t Z_t \right| dt$$

$$\leq \int_0^{t_0} \left| N^\alpha s_N \left(\frac{\tilde{W}_t}{N^\beta} \frac{\tilde{Z}_t}{N} \right) - s W_t Z_t \right| + \left| N^\alpha s_N \frac{\tilde{W}_t}{N^\beta} \frac{\tilde{Y}_t}{N} \right| + \left| N^\alpha r_N \frac{\tilde{X}_t}{N} \frac{\tilde{Y}_t}{N^\beta} - r X_t Y_t \right| +$$
$$\left| N^\alpha r_N \frac{\tilde{W}_t}{N} \frac{\tilde{Z}_t}{N^\beta} - r W_t Z_t \right| dt$$

$$= \int_0^{t_0} \left(|N^\alpha s_N - s| \, W_t Z_t + \left| N^\alpha s_N N^{\beta-1} W_t Y_t \right| + |N^\alpha r_N - r| \, X_t Y_t + |N^\alpha r_N - r| \, W_t Z_t \right) dt$$

Now, for some constant $K > 0$, we stop the process when either W_t or Y_t reaches K. Suppose this time is T_K. Then, as $N \to \infty$,

$$\int_0^{t_0 \wedge T_K} \left(|N^\alpha s_N - s| \, W_t Z_t + \left| N^\alpha s_N N^{\beta-1} W_t Y_t \right| + |N^\alpha r_N - r| \, X_t Y_t + |N^\alpha r_N - r| \, W_t Z_t \right) dt \to 0$$

Using similar calculations, we get

$$\int_0^{t_0 \wedge T_K} \left| \beta_{X,N}(\tilde{W}_t, \tilde{X}_t, \tilde{Y}_t, \tilde{Z}_t) - b_x(W_t, X_t, Y_t, Z_t) \right| dt \to 0$$

$$\int_0^{t_0 \wedge T_K} \left| \beta_{Y,N}(\tilde{W}_t, \tilde{X}_t, \tilde{Y}_t, \tilde{Z}_t) - b_y(W_t, X_t, Y_t, Z_t) \right| dt \to 0$$

$$\int_0^{t_0 \wedge T_K} \left| \beta_{Z,N}(\tilde{W}_t, \tilde{X}_t, \tilde{Y}_t, \tilde{Z}_t) - b_z(W_t, X_t, Y_t, Z_t) \right| dt \to 0$$

as $N \to \infty$. By taking K sufficiently large, we see that $P(\Omega_1) \to 1$ as $N \to \infty$ and so $P(\Omega_1^c) = 0$.

Lastly, suppose A is small. Let C_0, C_1, C_2 be constants. Then,

$$\int_0^{t_0} \alpha(W_t) dt = \int_0^{t_0} \left(\sum_{W_t' \neq W_t} |W_t' - W_t| q((W_t, X_t, Y_t, Z_t), (W_t', X_t', Y_t', Z_t')) dt \right)$$
$$\leq \int_0^{t_0 \wedge T_K} C_0 \cdot \frac{1}{N^{2\beta}} (N^\alpha N(1 + r_N)) dt$$
$$\leq \int_0^{t_0 \wedge T_K} C_0 \cdot \frac{1}{N^{2\beta}} (C_1 N^{1+\alpha}) dt$$
$$\leq \int_0^{t_0 \wedge T_K} C_2 N^{1+\alpha-2\beta} dt \to 0$$

as $N \to \infty$ (since $\beta > (1+\alpha)/2$). So for sufficiently large N, $P(\int_0^{t_0} \alpha W_t dt \leq A t_0) = 1$. Similarly, for sufficiently large N, $P(\int_0^{t_0} \alpha X_t dt \leq A t_0) = 1$, $P(\int_0^{t_0} \alpha Y_t dt \leq A t_0) = 1$, and $P(\int_0^{t_0} \alpha Z_t dt \leq A t_0) = 1$. So $P(\Omega_2) = 1$ and hence $P(\Omega_2^c) = 0$.

Thus $P(\Omega_0^c \cup \Omega_1^c \cup \Omega_2^c) \to 0$ as $N \to \infty$, and thus

$$P\left(\sup_{t \leq t_0} |W_t - w_t| > \epsilon \right) \leq 4 A t_0 / \delta^2$$

$$P\left(\sup_{t\leq t_0}|X_t - x_t| > \epsilon\right) \leq 4At_0/\delta^2$$

$$P\left(\sup_{t\leq t_0}|Y_t - y_t| > \epsilon\right) \leq 4At_0/\delta^2$$

$$P\left(\sup_{t\leq t_0}|Z_t - z_t| > \epsilon\right) \leq 4At_0/\delta^2.$$

And since A can be arbitrarily small, then for all $\eta > 0$, we have

$$P\left(\sup_{t\leq t_0}|W_t - w_t| > \epsilon\right) \leq \eta$$

$$P\left(\sup_{t\leq t_0}|X_t - x_t| > \epsilon\right) \leq \eta$$

$$P\left(\sup_{t\leq t_0}|Y_t - y_t| > \epsilon\right) \leq \eta$$

$$P\left(\sup_{t\leq t_0}|Z_t - z_t| > \epsilon\right) \leq \eta.$$

□

Now, we get that $x_t + z_t = 1$ because the number of individuals of the other two types is of order $O(N^\beta)$, and $N^\beta \ll N$. The equations for x_t and z_t can be solved to give $x = L_t$ and $z_t = 1 - L_t$, where L_t is the function in (1). So, we can then eliminate the variables x and z to get three equations:

$$\begin{aligned} w'(t) &= sw(1-L) + r(yL - w(1-L)) \\ &= sw(1-L) + ryL - rw + rwL \\ &= sw - swL - rw - rwL + rLy \\ &= (s-r)(1-L)w + rLy \end{aligned}$$

Doing similar calculations for y, we get:

$$\begin{aligned} y'(t) &= -syL + r(w+y)(1-L) - ry \\ &= -syL + rw - rwL - ryL \\ &= r(1-L)w - (r+s)Ly \end{aligned}$$

and

$$L'(t) = sL(1-L)$$

is a logistic function.

This leaves us with a system of three differential equations:

$$w'(t) = (s-r)(1-L)w + rLy$$

$$y'(t) = rw(1-L) - yL(r+s)$$

$$L'(t) = sL(1 - L)$$

These equations model the rate of change of the number of individuals with the beneficial allele at the selected and neutral sites at each given point in time. The differential equation for w shows the rate of change of the number of alleles with the type-0 allele at the selected site and the type-0 allele at the neutral site as a function of time, and the differential equation for y shows the rate of change of the number of individuals with the type-0 allele at the selected site and the type-1 allele at the neutral site as a function of time.

Therefore, the combined quantity $w' + y'$ shows the rate of change of the number of individuals with the type-0 allele at the neutral site as a function of time. We also know that the logistic equation L shows the rate of change of the number of individuals with the type-0 allele at the selected site over time.

By solving these differential equations, we can model the number of individuals with the type-0 allele at the selected site and the neutral site, separately. We have not found a closed-form solution to these equations.

3 Beginning and End of the Sweep

In this section, we show that in the two locus case, recombination is unlikely to occur during the beginning and end of the sweep. This would imply that most of the recombination occurs during the middle of the sweep, where the above differential equation approximations can be applied.

The results described in the coming sections follow from proofs described in Schweinsberg and Durrett [6].

3.1 Beginning of the Sweep

We create a separate model for the beginning and end of the selective sweep. For these two ends, we switch to thinking backwards in time when doing our calculations. We first model the beginning of the sweep in this section.

We state that a lineage is in the type-0 population at time t if it has the beneficial allele in the neutral site at time t. Otherwise, we state that the lineage is in the type-1 population at time t. Now, we randomly sample a lineage at the end of the selective sweep, or when the number of type-0 individuals at the selected site reaches N, and try to keep track of its ancestry. We want to prove that if its descended from the original type-0 individual, then it is highly likely that it is also in the type-0 population at a time close to time 0.

In mathematical terms, for all $j \in \{0, ..., N\}$, we let τ_J be the most recent time at which the number of type-0 individuals is equal to J (i.e., looking backwards in time, the first time in which the number of individuals is equal to J). Then, if a lineage is in the type-0 population at the start of the sweep, then we try to show that with high probability, the lineage is in the type-0 population at time τ_J.

Suppose that at time τ_J, a lineage is in the type-0 population. Here, we use a result from Schweinsberg and Durrett [6] showing that the probability that a lineage recombines from the type-0 to the type-1 population when there are k type-0 individuals is approximately $1 - \exp\{-r/ks\}$. Then, the probability that no recombination occurs at any step is

$$P(\text{no recombination}) \approx \exp\left(-\sum_{k=1}^{N} \frac{r}{ks}\right) \asymp \exp\left(-\frac{r}{s}\log N\right) = N^{-r/s}.$$

Now, suppose that the lineage starts in the type-1 population, recombines before time τ_J, and does not recombine after that. By Proposition 2.2 in Schweinsberg and Durrett [6], we know that the probability that the lineage recombines from type-1 to type-0 when there are k individuals with the beneficial allele is $r/(s(N-k))$. And, the probability of no recombination after that is $\exp\{-(r/s)\log k\}$. So, this probability is approximately equal to:

$$P(\text{recombination when } k \text{ type-0 individuals}) \approx \frac{r}{s(N-k)} \exp\left(-\frac{r}{s}\log k\right)$$

So the total probability of this occurring, accounting for all values of k, is

$$\begin{aligned} P(\text{type 0 at time 0} \mid \text{type 1 at time } \tau_J) &\approx \sum_{k=1}^{J} \frac{r}{s(N-k)} \exp\left(-\frac{r}{s}\log k\right) \\ &\leq C\frac{r}{sN}\sum_{k=1}^{J} k^{-r/s} \\ &\leq C\frac{r}{sN} J^{(1-r/s)}. \end{aligned}$$

for some constant C. Now, suppose $J = \delta N$, where $\delta \ll 1$. Then,

$$P(\text{type 0 at time 0} \mid \text{type 1 at time } \tau_J) \approx \frac{r}{sN}(\delta N)^{(1-r/s)} = \frac{r}{s}\delta^{(1-r/s)}N^{-r/s} \ll N^{-r/s}$$

as $\delta \ll 1$ and $r < s$. Therefore the probability of one recombination occurring is significantly less than the probability of no recombinations occurring, and therefore, $P(\text{type 0 at time } \tau_J \mid \text{type 0 at time 0})$ goes to 1.

This shows that for $J = \delta N$ where $\delta \ll 1$, the probability that any given lineage descended from the original beneficial allele did not recombine into the ancestral population before time τ_J is high, and that no recombination before time τ_J is the most likely outcome for this lineage. Therefore the probability that recombinations occur early in the sweep is low. We want to find the probability that a lineage is descended from the individual that got the mutation in the neutral site, and we show that the probability is not affected much by recombination early on in the sweep.

Therefore we can ignore the recombinations that occur at the beginning of the sweep, as they are negligible compared to those in the middle of the sweep. Most of the recombinations occur during the middle of the sweep, which the previous differential equation approximation model can be applied to.

3.2 End of the Sweep

Towards the end of the sweep, we use the same logic as at the beginning of the sweep, except that we examine time τ_{N-J} instead. We aim to show that if a lineage has the type-0 allele at the neutral site at the end of the sweep, it also has the type-0 allele at the neutral at time τ_{N-J}.

In this case, we examine the model backwards in time. We let $R(i)$ be the time of the first recombination going backwards in time. We use a method similar to the ones in Propositions 2.2 and 2.1 in Schweinsberg and Durrett [6] to get the result:

$$
\begin{aligned}
&P(R(i) \geq \tau_{N-J}) \\
&\leq \frac{Cr^2}{sN(N-J-1)} + \frac{Cr}{s(N-J)} + \frac{Cr}{s\sqrt{N-J}} + \left(1 - \left(\frac{N-J}{N}\right)^{r/s}\right) + O\left(\frac{1}{N} + \frac{(1-s)^{N-J}}{(N-J)\log N}\right) \\
&\leq \frac{C}{N(2N-J-1)} + \frac{C}{N-J} + \frac{C}{\sqrt{N-J}} + \left(1 - \left(\frac{N-J}{N}\right)^{r/s}\right) + O\left(\frac{1}{N} + \frac{(1-s)^{N-J}}{(N-J)\log N}\right) \\
&\to 0 \text{ (as } N \to \infty)
\end{aligned}
$$

So for all fixed J, as N gets large, $P(R(i) \geq \tau_{N-J}) \to 0$ as $N \to \infty$. And so the probability that there is at least 1 recombination after time τ_{N-J} is 0. Hence, if a lineage is in the type-0 population at the end of the sweep, it is highly likely that it is in the type-0 population at time τ_{N-J}.

This combined with the previous section show that recombinations are unlikely to occur towards the beginning and end of the selective sweep.

4 References

[1] Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. (2005), Genomic scans for selective sweeps using SNP data. Genome Res. 15(11):1566-75. doi: 10.1101/gr.4252305.

[2] Moran, P. A. P. (1958). "Random processes in genetics". Mathematical Proceedings of the Cambridge Philosophical Society. 54 (1): 60–71. doi:10.1017/S0305004100033193.

[3] Achaz, G., Schertzer, E. (2023), "Weak genetic draft and the Lewotin's paradox," bioRxiv preprint,
doi: https://doi.org/10.1101/2023.07.19.549703

[4] Darling, R.W.R., Norris, J.R., (2008), "Differential equation approximations for Markov chains". Probability surveys. 5: 37-79. doi: http://dx.doi.org/10.1214/07-PS121

[5] Kurtz, T.G., (1981), "Approximation of Population Processes". CBMS-NSF Regional Conference Series in Applied Mathematics. doi: https://doi.org/10.1137/1.9781611970333.ch1

[6] Schweinsberg, J., Durrett, R. (2005), "Random partitions approximating the coalescence of lineages during a selective sweep," The Annals of Applied Probability, vol. 15, no. 3,
doi: https://doi.org/10.1214/105051605000000430.