September 13, 2020

Abstract

My research focuses on the developement of Monte Carlo methods for phylogenetic inference of DNA sequences, in particular on the developement of Sequential Monte Carlo (SMC) algorithms for online onference of the spread of diseases. In this thesis I will explain the work on phylogenetic inference, and the results I have had after developing statistical algorithms using a software package named BEAST (the acronym stands for Bayesian Evolutionary Analysis by Sampling Trees). I will explain the benefits and limitations of several Monte Carlo methods, starting from Importance Sampling (IS) , to the Annealed Importance Sampling (AIS) and finally Sequential Monte Carlo (SMC), showing that the latter presents greater efficacy in complex problems.

0.1 Introduction

0.2 Notation

In the following chapters, unless otherwise stated, we will use the symbols as follow:

- the capital letters, for example P, for probability distributions
- the lowercase letters p, g for probability density functions
- θ and x to indicate the random variables to estimate
- y for the observations

0.3 Bayesian Statistics

Given a **probability space** $(\Omega, \mathcal{F}, P)$, a **random variable** X is a measurable function $X : \Omega \to E$, with E a measurable space, s.t. the probability of $X(\omega) \in S \subseteq E$, $\omega \in \Omega$, is given by $P(X^{-1}(S))$. Bayes rule allows the expression of conditional probability of random variables. If we consider two random variables, A and B, we express their joint probability, as follows

$$P(A \cap B) = P(A|B)P(B) \tag{1}$$

From (1), and using the obvious equivalence $P(A \cap B) = P(B \cap A)$, we can express the conditional probability of A given B as the product of the probability of A times the probability of B given A, as follows

$$P(A|B) = P(B|A)P(B) \tag{2}$$

Stated in the terms of statistical modelling, the problem we are interested in is defining the distribution of some parameters of a model that we want to estimate

$$\theta = [\theta_1, \theta_2, ..., \theta_D]^T \tag{3}$$

given a set of observations

$$y = [y_1, y_2, ..., y_N]^T \tag{4}$$

and in this context we express the formula (2) as

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)d\theta} \tag{5}$$

in equation (5), we define:

- **marginal likelihood** the term p(y) in the denominator
- **posterior** the term $p(\theta|y)$ on the left hand side of the equation
- **prior** the term $p(\theta)$
- **likelihood** the term $p(y|\theta)$

The prior can be considered as the a-priori information we have on the distribution of the parameters, the likelihood is the information that allows us to update the model after the data y that we observe.

0.4 The Monte Carlo method

In Bayesian statistics we often have the need to calculate expectations with respect to the posterior density $p(\theta|y)$ defined in (5), and there are cases where $p(\theta|y)$ cannot be computed in closed form, and might also be computationally non-tractable with traditional numerical integration methods (that suffer from the so-called *curse of dimensionality* as the dimension of the state space increases and the convergence rate can become exponentially worse [27]), and therefore we might not be able to evaluate the quantity

$$E_{p(\theta|y)}(f(\theta)) = \int f(\theta)p(\theta|y)d\theta = \mu \tag{6}$$

We will, in this section, explain the foundations of **Monte Carlo methods** [3] [15]. If we are able to draw N independent samples $\theta_1, \theta_2 ... \theta_N$ from $p(\theta|y)$, we can consider the approximation

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} f(\theta_i) \tag{7}$$

The quantity $\hat{\mu}$ of (7) approximates the expectation

$$\frac{1}{N}\sum_{i=1}^{N} f(\theta_i) \approx \int f(\theta)p(\theta|y)d\theta \tag{8}$$

The left-hand-side sum of equation (8) equals the right-hand-side integral almost surely in the limit $N \to \infty$ by the **strong law of large numbers** (note that in the left-hand-side the $p(\theta|y)$ is implicitly approximated by the random measure since we are drawing the θ_i samples from it). We will explore in the following sections of this chapter how to deal with cases when we are not able to draw the samples $\theta_1, \theta_2 ... \theta_N$ directly from the posterior distribution $p(\theta|y)$, and we will see how to write an approximation conceptually similar to equation (7). By simple application of the linearity of the expectation we have that the estimator (7) is unbiased, i.e. that

$$E_{p(\theta|y)}(\hat{\mu}) = E_{p(\theta|y)}(f) \tag{9}$$

and that the variance of the estimator is given by

$$var(\hat{\mu}) = var(\frac{1}{N}\sum_{i=1}^{N} f(\theta_i)) = \frac{\sigma^2}{N} \tag{10}$$

assuming that the variance of f

$$\sigma^2 = E_{p(\theta|y)}(f(\theta)^2) - \mu \tag{11}$$

be finite.

0.5 Importance Sampling

Importance sampling (IS) provides a way to estimate integrals by the use of an instrumental auxiliary distribution. In detail, let's assume that we want to calculate the expectation of a function f according to the density p, assumed to be 0 outside $\mathbf{D} \subseteq \mathbb{R}^n$

$$E_p(f(\theta)) = \int_{\mathbf{D}} f(\theta)p(\theta)dx = \mu \tag{12}$$

we may want to approximate the integral via use of the sum as in equation (7), but let's assume we cannot draw easily samples from the distribution

p. We can use an auxiliary distribution, defined **proposal distribution**, g, easier to draw from: it is sufficient that $g(\theta) > 0$, $\forall \theta \in \mathbf{Q}$, where $\mathbf{Q} = \{\theta | f(\theta)p(\theta) \neq 0\}$. The validity of the process is shown by the equivalences below [16]

$$\begin{aligned} E_g(\frac{f(\theta)p(\theta)}{g(\Theta)}) = \int_{\mathbf{Q}} \frac{f(\theta)p(\theta)}{g(\theta)} g(\theta)d\theta = \int_{\mathbf{Q}} f(\theta)p(\theta)d\theta \\ = \int_{\mathbf{D}} f(\theta)p(\theta)d\theta + \int_{\mathbf{D}^c \cap \mathbf{Q}} f(\theta)p(\theta)d\theta - \int_{\mathbf{Q}^c \cap \mathbf{D}} f(\theta)p(\theta)d\theta = \mu \end{aligned} \tag{13}$$

The last equality comes from $p(\theta) = 0$ in $\mathbf{D}^c$ and $f(\theta) = 0$ in $\mathbf{Q}^c$. Therefore the importance sampling estimate becomes

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(\theta_i)}{g(\theta_i)} f(\theta_i) = \frac{1}{N} \sum_{i=1}^{N} w(\theta_i) f(\theta_i),\ \theta \sim g \tag{14}$$

Where, in equation (14), the so-called weights are defined as follows

$$w(\theta) = \frac{p(\theta)}{g(\theta)} \tag{15}$$

The weights w correct for the fact that we are sampling from the proposal distribution, g, instead of p. Therefore, we can sample $\theta_1, \theta_2, ..., \theta_N$ independently from the proposal distribution g, and due to the LLN (as in formula (7)) we have

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} w(\theta_i) f(\theta_i) \xrightarrow[N \to \infty]{a.s.} E_p(f(\theta)) \tag{16}$$

Similarly to equation (9) (see also [4] and [16]), it is easy to demonstrate that μ of equation (14) is an unbiased estimator of the mean, i.e. that

$$E_g(\hat{\mu}) = \mu \tag{17}$$

and that [16]

$$Var_g(\hat{\mu}) = \frac{\sigma^2}{N} \tag{18}$$

where

$$\sigma^2 = \int_{\mathbf{Q}} \frac{\left(f(\theta)p(\theta)\right)^2}{g(\theta)} d\theta - \mu^2 = \int_{\mathbf{Q}} \frac{\left(f(\theta)p(\theta) - \mu g(\theta)\right)^2}{g(\theta)} d\theta \tag{19}$$

The formulae (19) give us a way to analyse an optimal proposal g: from the second expression in (19) we see that an optimal proposal will minimise the numerator $(f(\theta)p(\theta) - \mu g(\theta))$, therefore a function g proportional to fp, and ideally [16]:

$$g_{opt} = \frac{|f|p}{E_p(|f|)} \tag{20}$$

although the g_{opt} of equation (20) is not practically feasible because it would mean that we can sample directly from p (which by assumption is not the case). It is therefore advisable in a good proposal choice that g is proportional to $|f|p$ (for example it has spikes where $|f|p$ does) [16]. We can also see from the second expression in (19) that small values of g in the denominator would magnify whatever lack of proportionality in the numerator between g and $|f|p$ [16], therefore we want a proposal that has heavier tails than p (or at least as heavy as p) [4].

0.5.1 Estimating the normalization constant through IS

In Bayesian analysis we can usually only compute an un-normalised version of p, or g or both. For example, in the case of p, we may have $p = Z\hat{p}$, where $\hat{p}$ is the normalised distribution and Z is the normalising constant. Let's assume without loss of generality that only p is unnormalised, we have therefore that

$$\int_{\Theta} p d\theta = Z \tag{21}$$

We can use IS to estimate the normalising constant by considering that, from equation (21), we have

$$E_g(w(\theta)) = \int \frac{p(\theta)}{g(\theta)} g(\theta) d\theta = \int p(\theta) d\theta = Z \tag{22}$$

Therefore using again formulas of (14) and (16) applied to equation (22), we have that

$$\hat{Z} = \frac{1}{N} \sum_{i=1}^{N} w_i \tag{23}$$

The $\hat{Z}$ of equation (23) is the estimate of the normalising constant Z up to which we know the distribution p (a similar procedure can be applied if g

is known up to a normalising constant). Using the result of (23), we can write a "self-normalised" version of the estimate of equations (14) and (16), as follows

$$\hat{\mu} = \frac{\sum_{i=1}^{N} w(\theta_i) f(\theta_i)}{\sum_{i=1}^{N} w_i} \tag{24}$$

The normalised weights of equation (24) are such that

$$\hat{w_i} = \frac{w_i}{\sum_{j=1}^{N} w_j} \tag{25}$$

Formula (24) is the "self normalised" version of equations (14) and (16), with the property that the weights $\hat{w}^i$ of equation (24) add up to 1 (as seen from (22) and (23), this means that we are simulating drawing from normalised distributions). It is not difficult to show that (see for example [4])

$$E_g(\hat{\mu}) = \mu + \frac{\mu Var_g(w(\theta)) Cov_g(w(\theta), w(\theta) f(\theta))}{N} + O(N^2) \tag{26}$$

And that the variance is

$$Var_g(\hat{\mu}) = \frac{Var_g(w(\theta) f(\theta)) 2\mu Cov_g(w(\theta), w(\theta) f(\theta)) + \mu Var_g(w(\theta))}{N} + O(N^2) \tag{27}$$

We can therefore see from (26) that $\hat{\mu}$ is biased, it has though the advantage that it can be calculated knowing the density up to a constant, in fact the normalising constant cancels out in the calculation (as shown in [4]).

0.5.2 Measuring the performances of an IS estimator

As we have seen in the previous sections, the Importance Sampling method allows us to perform calculation of integrals, for example the expectation of a function f w.r.t. a density p, like in equation (12), by using samples drawn from a proposal distribution g, with a sum like in equation (14). An obvious question to ask is how does the IS approximation compare with the usual Monte Carlo approximation [18] [19] of the integral that we would have by drawing samples from the distribution p, as in (7). The remainder of this section will be dedicated to answering this question, using the logic outlined in [17]. For ease, we rewrite here some definitions used in the previous

sections, changing slightly the notation (to make it coherent with [17]). We start from the integral of an expectation

$$I = E_p(f) = \int f(\theta)p(\theta)d\theta \tag{28}$$

Then we call $\bar{I}$ the Monte Carlo approximation of (28)

$$\bar{I} = \frac{1}{N}\sum_{i=1}^{N} f(\theta_i),\ \theta \sim p \tag{29}$$

And we call $\tilde{I}$ the self normalised IS approximation of (28) (as in (24))

$$\tilde{I} = \sum_{i=1}^{N} \bar{W}(\theta^{(i)})f(\theta^{(i)}),\ \theta \sim g \tag{30}$$

Where, in equation (30), $\bar{W}$ are the self-normalised weights of equation (25). A measure that has been widely used [16] [17] to compare the performance of IS estimators, is the so-called **Effective Sample Size (ESS)**, which compares the variances of the traditional Monte Carlo estimate (29) and the IS approximation (30)

$$ESS = N\frac{Var(\bar{I})}{Var(\tilde{I})} \tag{31}$$

We can notice that the ESS of (31) has some drawbacks, for example it depends on the integrand function f (as clearly seen from (28) and (30)), and therefore an estimator that is good for an integrand function f_1 may not in general be good for another function f_2, and also in order to calculate (31) we will need to compute integrals that are in general intractable as the integral (28) that we are trying to estimate (see for example [17] for a detailed expression of such integrals). Therefore some simplifications have been used [17] [19] that reduce significantly the complexity of (31), to:

$$ESS \approx \frac{N}{1 + Var_g(W)} \tag{32}$$

We see from equation (32) that, in the ideal case where the weights are known exactly (and therefore with zero associated variance), we have $ESS = N$, i.e. we are in a situation that is as good as if we were drawing directly from

the target distribution. As we see in [17] [18] [19], further simplification of (32) bring to

$$ESS \approx \frac{NZ^2}{E_g(\bar{W}^2)} \tag{33}$$

Where Z is the normalising constant expressed in (21). It is to be noted that (33) has, as said, approximations and that these restrict the validity of (33) to cases where the approximations are valid [19] [17] (for example since there is no more dependance on the integrand function f, it is assumed that the proposal g is "reasonably" close to the optimal proposal (20)). By using particle approximations for Z from (23) and $E_g(\bar{W}^2) \approx \frac{1}{N}\sum_{j=i}^{N}(w^{(j)})^2$ which brings us to the final approximation of ESS in the version widely used in literature:

$$\hat{ESS} = N\frac{(\sum_{j=i}^{N} w^{(j)})^2}{\sum_{j=i}^{N}(w^{(j)})^2} = \frac{1}{\sum_{j=i}^{N}(\bar{w}^{(j)})^2} \tag{34}$$

Where, in (34), in the first equation the $w^{(j)}$ are unnormalised weights of equation (15), whereas in the second equation the $\bar{w}^{(j)}$ are the self-normalised weights of equation (25).

0.6 Markov Chain Monte Carlo (MCMC)

We will introduce in this section Markov Chains Monte Carlo (MCMC). Like the Monte Carlo methods in previous sections, MCMC provides an indirect way to approximate drawing samples from a distribution that we cannot directly draw from. Unlike the traditional MC and Importance Sampling, MCMC samples are not independently distributed: as we can see from equation (35) in fact, there is conditional dependence between values. A stochastic process is defined Markov Chain Monte Carlo if, considering $X_1, X_2, ...X_N$ random variables defined on a common probability space $(\chi, \mathcal{A}, P)$, that are the realization of the process, it has the following so-called Markov property:

$$P(X^{(t)} = x^{(t)}|X^{(t-1)} = x^{(t-1)}, ..., X^{(1)} = x^{(1)}) = P(X^{(t)} = x^{(t)}|X^{(t-1)} = x^{(t-1)}) \tag{35}$$

As we can see from (35), the value of the chain at a particular time t is only dependent from the value at time $t-1$. In the following parts of this section we will outline the basic concepts that will help us introduce MCMC [15].

0.6.1 Markov Kernel

Considering two measurable spaces (χ, A), (Y, B) a transition kernel is a map [15] $K : \chi \times B \to [0, 1]$, s.t.

- $\forall x \in \chi$, $k(x, \cdot)$ is a probability measure
- $\forall B_i \in B$, $k(\cdot, B_i)$ is measurable

The kernel is a conditional probability density, we will speak more extensively about the associated probability measure in the next subsections. In the continuous case we have that

$$P(X_t \in B_i | X_{t-1} = x_{t-1}) = \int_{B_i} k(x_{t-1}, x_t) dx_t \tag{36}$$

whereas in the discrete case we have

$$P(X_t \in B_i | X_{t-1} = x_{t-1}) = \sum_{x' \in B_i} k(x_{t-1}, x') \tag{37}$$

0.6.2 Initial distribution of the chain

The chain X_n is defined for $n \in \mathbb{N}$, therefore there is a X_0, starting point of the chain. We define the initial distribution of X_0 as μ

$$P(X_0 \in A) = \int_A \mu(x) dx \tag{38}$$

And, consequently

$$P(X_0 = x_0) = \mu(x_0) \tag{39}$$

The extension to the discrete case of (38) is similar to formula (37), we have in fact

$$P(X_0 \in A) = \sum_{x_{0i} \in A} \mu(x_{0i}) \tag{40}$$

0.6.3 Joint and conditional distributions of the X_i

Continuing what we have said in the previous section 0.6.2, the marginal distribution of X_1 is obtained by $\mu_1(x_1) = \mu(x_0)k(x_0, x_1)$, and consequently [15]

$$P(X_1 \in A | X_0 = x_0) = \int_A \mu(x_0) k(x_0, x_1) dx_1 \tag{41}$$

and, finally

$$P(X_1 \in A_1, X_0 \in A_0) = \int_{A_0} \int_{A_1} \mu(x_0) k(x_0, x_1) dx_0 dx_1 \tag{42}$$

By using the Markovian property (35) and the definition of the kernel we have that [4]

$$P(X_0 = x_0, ...X_n = x_n) = \mu(x_0) \prod_{j=1}^{n} k(x_{j-1}, x_j) \tag{43}$$

We also introduce the notation used in literature [15] [4] $k^1(x, A) = k(x, A)$, and

$$k^s(x_t, x_{t+s}) = \int_{A^{s-1}} \prod_{j=t+1}^{t+s} k(x_{j-1}, x_j) dx_t ... dx_{t+s-1} \tag{44}$$

so we can express (43) as

$$P(X_0 = x_0, ...X_n = x_n) = \mu(x_0) k^n(x_0, x_n) \tag{45}$$

0.6.4 Stationary property of the MCMC and invariant distribution

We will in this section introduce the concept of **stationary/invariant distribution** of the chain, i.e. a distribution π s.t. [15]:

$$X_n \sim \pi \rightarrow X_{n+1} \sim \pi \tag{46}$$

A σ-finite measure π is invariant for the kernel $k(\cdot, \cdot)$ and the chain X_N if

$$\pi(A) = \int_{\chi} k(x, A) \pi(x) dx \tag{47}$$

Equation (47) states the condition (46) (if the invariant distribution is a probability measure it is also called *stationary* due to (46)). A theorem states [15] that if X_N is a **recurrent** chain then it has an invariant σ-finite measure which is unique up to a constant (*recurrence* is a property that states that, whatever the initial condition of the chain, we will end up in a set A having positive measure an infinite number of time as $N \rightarrow \infty$ [15]). The stationary property of the MCMC can also be related to another property, that states

that states that the direction of time does not matter in the dynamic of the chain, $P_{X_{n+1}}(X_{n+1}|X_n = x) = P_{X_n}(X_n|X_{n+1} = x)$. This property is called **reversibility** and is stated as follows [4] [15]:

$$k(x, y)f(x) = k(y, x)f(y) \tag{48}$$

Equation (48) is named *detailed balance condition* and provides a sufficient condition for f to be a stationary distribution for the chain. It is easy to prove that [15] if a transition kernel k satisfies the condition (48), with π a probability density function then π is the invariant density of the chain.

0.6.5 Convergence of the MCMC

We are interested in understanding if and towards what the chain X_n is converging. We have seen in section 0.6.4 the conditions for existence of a stationary distribution for MCMC, in this section we will state the conditions for such stationary distribution to be the limiting distribution of the chain. We will, in this section, state the two convergence theorems of the chain to the stationary distribution: the **convergence by LLN** and, under stronger conditions, the **convergence in total variation norm**.

Convergence by LLN

Under the following conditions [15]: *if the chain is Harris-recurrent, with invariant measure π, then the following convergence theorem can be proved*

$$\lim_{n\to\infty} \frac{1}{n} \sum f(X_i) = \pi(f), \forall X_0, \text{ a.s. by LLN} \tag{49}$$

(*Harris-recurrence* is a stronger property than that of recurrence mentioned earlier in the section, the main concept anyway remains that whatever the initial condition of the chain, we will end up in a set A having positive measure an infinite number of time as $N \to \infty$ [15]). Formula (49) states that, given the conditions stated, no matter the initial condition X_0, the chain will converge by Law of Large Numbers to the stationary distribution π, for all π-integrable functions.

Ergodicity and convergence in total variation norm

We start by defining an additional property of the chain which will be auxiliary in the formulation of the convergence. We define **periodic** a chain X_n

which cyclically returns in the same states, mathematically, X_n is periodic with period d if there are non empty disjoint sets $A_0 ... A_{d-1}$ s.t.

$$K(x, A_j) = 1, \text{ for } j = i + 1(\text{mod d}), \forall i \text{ s.t. } x \in A_i \tag{50}$$

If the conditions of (50) are not met, then X_n is aperiodic. We can now state the following conditions of convergence [15]: *if the chain is Harris-recurrent, aperiodic with invariant measure π, then there is convergence of the chain to the stationary distribution π, whatever the initial distribution μ*:

$$\lim_{n\to\infty} \left|\left| \int_A k^n(x, \cdot)\mu(x) - \pi \right|\right|_{TV} \tag{51}$$

Where k^n is the transition kernel applied n times introduced in (45), and the *total variation norm* is

$$||\mu_1(A) - \mu_2(A)||_{TV} = sup_A|\mu_1(A) - \mu_2(A)| \tag{52}$$

Formula (51) states **ergodicity**.

0.6.6 Metropolis-Hastings algorithm

We now need a way to take advantage of the properties of the chain outlined in the previous sections, and build chains that converge to a stationary distribution of our choice: the **Metropolis-Hastings algorithm** [20] [21] has been built with this purpose. Suppose we want to draw samples from a target distribution π and we want to do so via a MCMC. The algorithm allows us to approximate drawing samples from an arbitrary distribution p. We build the transition kernel k in three steps in the following way:

1. draw samples from the proposal g: $X_n^* \sim g(X^*|X_n)$
2. calculate the *acceptance ratio* $\alpha(X_n^*|X_n) = \frac{p(X_n^*)g(X_n|X_n^*)}{p(X_n)g(X_n^*|X_n)}$
3. we draw from the uniform distribution $u \sim \text{Unif}[0, 1]$, if $u \leq \alpha$ we have $X_{n+1} = X_{n+1}^*$, otherwise $X_{n+1} = X_n$

The kernel k built with the three-step procedure outlined above can be synthesised as follows:

$$k(x_{n+1}, x_n) = \alpha(x_{n+1}^*|x_n)g(x_{n+1}^*|x_n) + \mathbf{1}_{\{x_{n+1}^* = x_n\}} \left[1 - \int \alpha(s|x_n)g(s|x_n)ds\right] \tag{53}$$

It can be demonstrated [4] [15] that the kernel (53) satisfies the reversibility condition (48) and therefore, as explained in section 0.6.4, p is the invariant distribution of the chain.

0.7 Annealed Importance Sampling (AIS)

We will, in this section, introduce a Monte Carlo method named **Annealed importance Sampling (AIS)**. Like the other methods introduced so far, the AIS is used to approximating drawing from a target distribution, in particular the AIS can be seen as an enhancement [1] to the Importance Sampling and MCMC tecniques introduced in sections 0.5 and 0.6, and we will see later in this section that it combines the two (MCMC and IS). Annealed Importance Sampling allows us to move from an initial tractable distribution to a target distribution of interest, which is intractable or difficult to draw from.

Annealed Importance Sampling vs Importance Sampling

We have seen in section 0.5 that an important measure for IS is the Effective Sample Size (ESS), a quantity that provides us with a measure of how close the IS process is to an ideal situation where we can draw directly from the target distribution. Studies have shown that, in some reference scenarios [28], the number of Importance samples N has to be increased exponentially as the dimension of the state space increases, to keep the ESS at a predefined level. Therefore the approximation of the target given by the IS method in general deteriorates exponentially with the dimension. It has to be noted that the "curse of dimensionality" for IS, does not always automatically happen [28], as there may be cases where there is a diminishing response of the likelihood to perturbations in growing coordinate index, and in such case increasing the state space dimension has only a mild effect in the cost of the problem. Also, a relatively recent field of study focuses on *active subspaces* [28]: in many application areas where models are formulated in high dimensional state spaces, there is often a small subspace which captures most of the features of the system. It is the dimension of this subspace that mainly affects the cost of the problem, and in presence of these active subspaces, we expect as well that the number of particles to keep a predetermined ESS will be given mainly by the dimensionality of the subspace, which is relatively constant,

and will not therefore be subject to the curse of dimensionality.

Aside from the particular cases, IS is therefore affected negatively (at an exponential rate) by high dimensions. In contrast, it has been shown that [29] [30] sequential Monte Carlo methods like AIS and Sequential Monte Carlo (SMC, introduced later in this section), scale polynomially as the dimension d of the state space grows large. In particular it has been shown [29] [30] that it is possible to find an approximation s.t. the ESS remains at a predermined level at a cost $O(Nd^2)$, where N is the number of particles and d the dimension of the state space. In addition, it is possible to approximate expectations wrt any marginal, at the cost, uniformly in d, of $\frac{1}{\sqrt{N}}$ [29] [30].

Annealed Importance Sampling vs MCMC

AIS is most useful in cases of isolated modes, it works better, in general, that IS because for importance sampling to work well the variance of the weights should not vary too much, and, in case of isolated modes, some of the samples can different modes, which will be assigned different weights and the variation in weights due to this will be large if some important modes are found only rarely. AIS will also in general work better than MCMC in case of isolated modes, in fact when MCMC is used to sample from complex distributions it must usually proceed by making only small changes to the state variables [1], and this causes the chain to move between modes only rarely, and so to take a long time to reach equilibrium, and will exhibit high autocorrelations for functions of the state variables out to long time lags. The AIS, allowing to gradually approach the desired distribution of interest by making use of interpolating distributions, is an approach to avoid this.

The AIS algorithm

We proceed with the description of the algorithm. We state again that the aim of the AIS algorithm [1] is, as all other Monte Carlo methods, to approximate the drawing of samples from a target distribution of interest: let's call this normalised target distribution p_n, and f_n the associated un-normalised version (in the rest of the section we will use either or both p and f with indexes e.g. p_j and f_j for normalised and un-normalised distributions). AIS makes use of the methods of IS and MCMC, in fact it will use intermediate proposal distributions to draw from, and will make use of MCMC to move between the intermediate steps. The algorithm starts by sampling from an

initial proposal distribution that we call f_0 (or a said p_0), often a candidate for this proposal is the prior [1]. As said, the AIS moves from the initial distribution to the target f_n with intermediate target distributions f_j

$$f_j(x) = f_n(x)^{\beta(j)} f_0(x)^{1-\beta(j)}, \quad j = 1, 2, ..N, \quad 0 \leq \beta(j) \leq 1 \tag{54}$$

We start the algorithm with $\beta(j) = 0$ and therefore with f_0 and we arrive in the last step to the target f_n with $\beta(j) = 1$. In the particular case where we use the prior as starting distribution, expressing the posterior f_n as prior times likelihood as in the standard Bayesian set-up (5) we have:

$$f_n(x) = f_0(x)l(x) \tag{55}$$

where, in (55), f_0 is the prior, and l is the likelihood. In this case, using equation (55) in (54), we have that

$$f_j(x) = f_n(x)l(x)^{\beta(j)}, \quad j = 1, 2, ..N, \quad 0 \leq \beta(j) \leq 1 \tag{56}$$

If we indicate with p_i the normalized probability distribution associated with each f_i, the algorithm, as described in section 2 of [1], is as follows

$$\begin{aligned} &\text{step 0) } x_0 \sim p_0 \\ &\text{step 1) MCMC to move from } p_0 \text{ to } p_1, \; x_1 \sim p_1 \\ &\text{step 2) MCMC to move from } p_1 \text{ to } p_2, \; x_2 \sim p_2 \\ &\qquad ... \\ &\text{step n) MCMC to move from } p_{n-1} \text{ to } p_n, \; x_n \sim p_n \end{aligned} \tag{57}$$

Where, in (57), the MCMC moves of the intermediate steps are performed, as explained in section 0.6.1, using Markov kernels $k_i(x_{i-1}, x_i)$ (for example with Metropolis-Hastings 0.6.6). Explaining the algorithm more in detail:

0. in step 0 we draw x_0 from the starting proposal distribution, by assumption easy to draw from, for example the prior p_0

1. in step 1 we apply a Markov kernel $k(x_0, x_1)$ with target p_1 of (54), that allows us to move in the state space and, using the results seen in the MCMC section 0.6, this means that we approximate drawing x_1 from the distribution p_1

2. similarly to what we did in the previous step, we move towards p_2 and we approximate drawing x_2 from the distribution p_2

... ...

n. in the last step we approximate drawing a sample x_n from the target density p_n

The algorithm (57) produces samples $x_n^{(i)}$, i=1,2,... that are drawn from the target distribution p_n, with approximations that we will discuss in the remainder of the section. Like Importance Sampling, each particle $x_n^{(i)}$ has a weight that accounts for the fact that we are not directly drawing from the target distribution p_n, we will see that the expression of the weight of each particle is as follows (please note that the super index i indicating the particle has been omitted for brevity in the following formula for the f_j and, in addition, we are using f_j instead of p_j, this is possible because it can be shown [1] that normalising constants cancel out in ratios):

$$w^{(i)} = \frac{f_1(x_0)}{f_0(x_0)} \frac{f_2(x_1)}{f_1(x_1)} ... \frac{f_n(x_n)}{f_{n-1}(x_n)} \tag{58}$$

Before explaining how it is obtained mathematically, we can see from its expression that (58) is made up by products of importance weights: each factor $\frac{f_{j+1}(x_j)}{f_j(x_j)}$ is, as seen in (15), the ratio of the target over the proposal, in fact each f_j, by construction, is the proposal for the f_{j+1}, and these intermediate steps allow, compared to IS, a smoother transition from the proposal to the target, in fact, by tuning $\beta(j)$ of equation (54), it is possible to have proposals that are closer to the targets, allowing for greater efficiency of the intermediate IS steps. The validity of (58) can be shown [1] using the results we have already obtained in section 0.5 and 0.6. In fact, if we consider an extended state space $(x_0, ...x_n)$, with a joint distribution:

$$f(x_0, ...x_n) = f_n(x_n)\tilde{k}_{n-1}(x_n, x_{n-1})...\tilde{k}_0(x_1, x_0) \tag{59}$$

We have that the marginal for x_n of equation (59) is the density we are looking to draw from (the target distribution). The $\tilde{k}_j$ are backward transition kernels associated to the MCMC moves of (57), and can be calculated by using the detailed balance condition of MCMC of (48):

$$\tilde{k}_j(x, x') = \frac{k_j(x', x)p(x')}{p(x)} \tag{60}$$

By rewriting equation (59) as

$$f(x_0, ...x_n) = f_n(x_n)\frac{f_{n-1}(x_n)}{f_{n-1}(x_n)}\tilde{k}_{n-1}(x_n, x_{n-1})...\frac{f_1(x_0)}{f_1(x_0)}\tilde{k}_0(x_1, x_0) = \\ k_{n-1}(x_{n-1}, x_n)\frac{f_n(x_n)}{f_{n-1}(x_n)}...k_0(x_0, x_1)\frac{f_1(x_0)}{f_0(x_0)}f_1(x_1) \tag{61}$$

If we take a look at the proposal distribution g of the procedure (57), starting from the first step $x_0 \sim p_0$ and considering all the subsequent applications of Markov kernels $k(x_j, x_{j+1})$, it has the form:

$$g(x_0, ...x_n) = f_0(x_0)k_0(x_0, x_1)...k_{n-1}(x_{n-1}, x_n) \tag{62}$$

Therefore, the AIS can be see as a multi-step importance sampling, and the expression of the weight for the whole importance sampling process, as seen in (15), is

$$w^{(i)} = \frac{f(x_0, ...x_n)}{g(x_0, ...x_n)} = \frac{f_1(x_0)}{f_0(x_0)}\frac{f_2(x_1)}{f_1(x_1)}...\frac{f_n(x_n)}{f_{n-1}(x_n)} \tag{63}$$

And (63) brings the result (58), which proves our case. Since, as we saw in the steps of (57), at each step $x_j \sim f_j$ and therefore the function f_j becomes the proposal for the next step $f_{j+1}(x_j)$, all the rules that apply to importance sampling choice of proposal hold (please see section 0.5). The choice of the proposal, as in importance sampling, is critical for the success of the algorithm, and we will see in later sections in the application to phylogenetic analysis of DNA sequences that in non-trivial cases smooth transitions between functions, i.e. small steps in the exponent β of formula (54), are needed to have an acceptable Effective Sample Size.

0.8 Sequential Monte Carlo (SMC)

Sequential Monte Carlo (SMC) methods are a collection of techniques used to approximate a target distribution.

SMC vs AIS

Like the AIS methods described in section 0.7, SMC uses IS and a sequence of proposals to approximate intermediate target functions, with the aim to create a set of particles that approximate a distribution of interest, not easy

to draw from. A difference of SMC wrt AIS lies in that SMC uses a technique called *resampling* to account for the fact that the variance of the particles weights usually increases with (algorithmic) time (in fact the AIS is a subcase of the SMC where we employ no resampling). A measure of the variability of weights commonly used is the ESS, as introduced already in sections 0.5 and 0.7. The resampling is a sampling with replacement from the current sample of particles using the weights and resetting them to 1/N afterwards (therefore if, for example, before the resampling a particle has a normalised weight twice as big as the others, it has twice as much chance than the others to "survive" the resampling process, and also to be "copied", i.e. resampled, into one or more new particles that will be a copy of it, whereas particles with little weight, and therefore little representative of the target, are more likely to be eliminated). A technique usually employed is to resample when the ESS drops below a given threshold (for example if the threshold is, say, 0.9, when the effective sample size drops below 90% of the number of particles N the resampling happens) [29].

The SMC algorithm

As said in the introduction, the **Sequential Monte Carlo (SMC)** algorithm will have many commonalities with the AIS seen in section 0.7. We will, similarly to section 0.7, make use of consecutive "neighbouring" distributions, i.e. distributions that are not too different one from another so that the proposals and the target distributions, at each step, are sufficiently close. The starting point is, as in the common Monte Carlo methods, that we are willing to draw samples from a target distribution π_n. We proceed through intermediate targets as in equation (54), and at each step the previous target become the proposal for the next target. We proceed in steps, similar to (57), we start by drawing from an initial distribution pi_0 easy to draw from, it can for example be the prior (in which case the expressions simplify as in equations (55) and (56)), and we go on constructing the first steps as done in (57). We present here the SMC version that makes use of resampling of the particles, we will explain further in the section what this implies. The steps of the SMC algorithm follow:

$$\begin{aligned} \text{step 0) } & x_0 \sim p_0 \\ \text{step 1) MCMC to move from } & p_0 \text{ to } p_1,\ x_1 \sim p_1 \end{aligned} \tag{64}$$

0. we draw $x_0 \sim p_0$ from the starting proposal distribution p_0 (for example

we could choose the prior), by assumption easy to draw from, and set the weights initially to $w_0^{(i)} = \frac{1}{N}$

1. we use the drawn particles as an importance sampler proposal (see section 0.5) for p_1 of equation (54), and we have a weight update of $w_1^{(i)} = w_0^{(i)} \frac{p_1(x_0)}{p_0(x_0)}$, this update reflects the weight of particles after the drawing process

2. we normalise the weights $w_1^{(i)}$ and we resample the particles according to their weight, so the bigger the normalised weight the more the particle will have a chance to be chosen in this resampling process: this resampling step allows us to eliminate particles where the proposal weakly represent the posterior, and will replicate particles where there is a strong representation of the posterior, all particles after resampling will again have weights of $w_0^{(i)} = \frac{1}{N}$

3. we wish to use the points of the state space obtained from the previous drawing done in step 0, for the next step. To do so, we need to move in the state space, so we apply a Markov kernel $k_1(x_0, x_1)$ with target p_1 of (54), that allows us to move from x_0 to x_1 in the state space and, using the results seen in the MCMC section 0.6, this means that we approximate drawing the points x_1 from the distribution p_1. The actual distribution of the drawn points after the MCMC step, from the theory (see section 0.6) is $\tilde{p}_1(x_1) = \int p_0(x_0)k_1(x_0, x_1)dx_1$, we update the weights with $w_1^{(i)} = w_0^{(i)} \frac{p_1(x_1)}{\tilde{p}_1(x_1)}$, and we have drawn $x_1 \sim k_1(x_0, \cdot)$

we see, from the last step in the above procedure, that the weight update in the last step is

$$w_1^{(i)} = w_0^{(i)} \frac{p_1(x_1)}{\int p_0(x_0)k_1(x_0, x_1)dx_0} \tag{65}$$

Since it is not easy, in general, to calculate $\int p_0(x_0)k(x_0, x_1)dx_1$, we rewrite the fraction in the RHS of (65) so that it can be expressed in non-integral form; we do so by writing, for some $L(x_1, x_0)$

$$p_1(x_1) = \int p_1(x_1)L_0(x_1, x_0)dx_0 \tag{66}$$

Where, in (66), the $L(x_1, x_0)$ is a backward kernel, built so that $p_1(x_1)$ is the x_1-marginal of the joint distribution $p_1(x_1)L(x_1, x_0)$. Therefore, instead of

marginalising, we write the contribution in the RHS of (65) using the joint distributions

$$w_1^{(i)} = w_0^{(i)} \frac{p_1(x_1) L_0(x_1, x_0)}{p_0(x_0) k_1(x_0, x_1)} \tag{67}$$

Since we can choose $L_0(x_1, x_0)$ of (67) at will (as long as equation (66) holds), we can choose L s.t.

$$p_1(x_1) L_0(x_1, x_0) = p_1(x_0) k(x_0, x_1) \tag{68}$$

and, substituting (68) in equation (65) we have

$$w_1^{(i)} = w_0^{(i)} \frac{p_1(x_0)}{p_0(x_0)} \tag{69}$$

By repeating the steps outlined above, we can normalise the weights of (69) as in step 2, we resample, then we move in the state space from x_1 to x_2 using a Markov kernel $k_2(x_1, x_2)$ having p_2 of (54) as target, and, with a procedure similar to the one that has brought us from equation (65) to (69), we have that

$$w_2^{(i)} = w_1^{(i)} \frac{p_2(x_1)}{p_1(x_1)} \tag{70}$$

0.9 Models of genetic evolution

The main goal of population genetics and of phylogenesis (we explain the differences of the two in section 0.9.2) is to infer the past history of populations and describe the evolutionary forces that have shaped their genetic variations. The current section will give a very brief and schematic explanation of the forces behind evolution.

0.9.1 DNA

DNA is packaged into **chromosomes**. Taking as example the human species, there are 46 chromosomes situated in the cells nuclei, 23 pairs, one of each chromosome is inherited by the mother, the other by the father. **Genes** are section of the chromosome situated in so-called **loci**, each gene is responsible for a trait, for example hair colour. Different variations of the same gene are called **alleles**. The expression of different alleles of the same gene will result in different characteristics, for example a different colour of hair, say brown or blonde.

0.9.2 Population genetics vs phylogenesis

As introduced at the beginning of this section 0.9, the forces that shape genetic evolution can vary, and the resulting differences in the DNA may be due to, just to shortlist some [22]:

- **random drift**: changes in frequency of combination of certain alleles in a population due to chance, resulting in change of frequency of specific traits in a population
- **mutations**: occasional errors in the replication of DNA
- **selection**: mutations that are more advantageous become more likely to be passed to the following generations

Population genetics usually deals with timescales within a generation [22], and data often consists of time series of allele frequencies, estimated from multiple DNA samples from the same generation, and the task is to estimate the changes of allele frequencies over time, this is done for example using the **Wright-Fisher model**, introduced in section 0.9.3.
Phylogenesis usually considers timescales of multiple generations [22], and data often consist of a single sample from each species, and the task is generally to infer such parameters as divergence times of the species and populations size. Such task is accomplished for example by the **Coalescent model**, introduced in section 0.10
Of course the above subdivision between population genetics and phylogenesis is to be taken as a reference and differences between the two can be blurred, for example in the cases where we consider data sets containing DNA sequences that comprise recently diverged species and we need to model both types of differences in the data: mutations that are still polymorphic (i.e. short timescale mutations where we still see expressions of all different alleles) and mutations that have been fixed as substitutions and have therefore a longer timescale[22].

0.9.3 The Wright-Fisher model

The WrightFisher model [23] [24] [22] accounts for the effects of the evolutionary forces on allele frequencies over time: random drift, mutation, selection. The model assumes a randomly mating population of finite constant size

reproducing in discrete generations, by allowing the individuals in generation $r+1$ to choose parents at random from the previous generation r. The model describes the stochastic behaviour through time of the frequency of an allele at a locus. We give here a short description of its most schematic version with a diploid (i.e. with two sets of chromosomes, one coming from each parent) **population of size** N which contains only two alleles, denoted AA and AB, and only subject to *random drift*. This is a reasonably good approximation for relatively short timescales.

Mathematical derivation with probability given by allele frequency [22]

The main concept of the model is related to allele frequency [22]. If we name $z(r)$ the number of individuals in a population at generation r that have a specific allele, say the allele AA, then the frequency of the allele AA is

$$x(r) = \frac{z(r)}{N} \tag{71}$$

The best guess on the number of alleles $z(r+1)$ in generation $r+1$ is based on the frequency of the allele in generation r, expressed in (71), and is a binomial with population N and probability $x(r)$, expressing the probability of having $z(r+1)$ successes

$$z(r+1)|z(r) \sim Bin(N, x(r)) \tag{72}$$

And therefore, expressing (72) with $x(r) = \frac{z(r)}{N}$

$$P([z(r+1)|z(r)] = k) = \binom{N}{k} \left(\frac{z(r)}{N}\right)^k \left(1 - \frac{z(r)}{N}\right)^{N-k} \tag{73}$$

By using the results for the binomial distribution, and remembering that $x(r)$ is proportional to $z(r)$ by the population size N, assumed to be constant, we have that mean and variance of the binomial (72) are as follows:

$$E[x(r+1)|x(r)] = x(r) \tag{74}$$

$$Var[x(r+1)|x(r)] = \frac{1}{N} x(r)(1 - x(r)) \tag{75}$$

We see that both equation (74) and (75) depend on the frequency of the allele in the previous generation. By iterating the two expressions we have that [22]:

$$E[x(r+1)|x(0)] = x(0) \tag{76}$$

$$Var[x(r+1)|x(0)] = x(0)(1-x(0))\left(1-\left(1-\frac{1}{N}\right)^r\right) \tag{77}$$

And, for big N we can use the approximation

$$Var[x(r+1)|x(0)] \approx x(0)(1-x(0))\left(1-\left(1-e^{-t}\right)\right) \tag{78}$$

with

$$t(r,N) = \frac{r}{N} \tag{79}$$

We can see from (79) that we can estimate the population size N only if the generation r is known, otherwise we can only have an estimate of the combined $t(r,N)$ of (79), that we can name *generation time*. We will see in the Coalescent method described in following sections that it is a common problem of population inference often to be able to estimate only a function of the number of generations and of the population size, not each of the two parameters separately, if no additional information is known [22] [12]. From equations (74) and (75) we can see that there are two equilibria:

1. $x(r) = 0$, when in a generation we reach zero number of individuals with the specific allele, this causes the expected value for the following generations to be zero as well, with zero variance

2. $x(r) = N$, i.e. all the individuals have the allele, and in the future generations all individuals will have the allele as well, with zero variance

The above conclusion brings us to say that, under the conditions of the model, if a certain alelle has small frequency, it is more likely to disappear after a few generations (it is more likely to reach the equilibrium $x(r+n) = 0$, for some n), whereas if it has a frequency close to 1 (nearly all the population has the allele), it is more likely to reach the equilibrium point $x(r+n) = N$, for some n, i.e. all the population will end up having the specific allele.

Mathematical derivation with probability given by the population size [25]

We report here also a slightly different mathematical derivation of the Wright-Fisher model [25], as we will use some the results in the following section 0.10 on Coalescent theory. We consider here the probability of having a certain number of allele at generation $r+1$ distributed according to a binomial, as in equation (72), but in this case instead of using the allele frequency of equation (71) as probability parameter of the binomial, we use

$$p = \frac{1}{N} \tag{80}$$

Therefore, we write an equation similar to (72), and, indicating with $z(r+1)$ the number of alleles in generation $r+1$ and using (80) we say that

$$z(r+1) \sim Bin(N, p) \tag{81}$$

Expressing (81)

$$P(z(r+1) = k) = \binom{N}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{N-k} \tag{82}$$

In the hypothesis of N large $P(z(r+1) = k)$ of equation (82) is almost Poisson distributed $z(r+1) \sim Po(1)$ and [25]

$$P(z(r+1) \approx \frac{e^{-1}}{k!} \tag{83}$$

Using (83), we see that the probability that a particular allele has no expression in the next generation is

$$P(z(r+1) = 0) \approx e^{-1} \approx 0.37 \tag{84}$$

And therefore, from the result of (84), the probability of expression of the allele is approximately

$$P(z(r+1) \neq 0) \approx 1 - e^{-1} \approx 0.63 \tag{85}$$

Extending the result of (85) at t generations in the future, considering the independence of the events, as per hypotheses of the Wright-Fisher model expressed at the beginning of section 0.9.3, we see that

$$P(z(r+t) \neq 0) \approx (1 - e^{-1})^t \approx (0.63)^t \tag{86}$$

And so, under the hypotheses of the model, after a few generations only a few lineages contribute to the current population [25], in fact, taking as an example a population size of $N = 10000$, after $t = 15$ generations, a number of approximately 10 lineages will have contributed to the current allele population

$$10000(0.63)^{15} \approx 10 \tag{87}$$

The remaining $10000 - 10 = 9990$ lineages that, in the example, were present 15 generations ago, have not survived.

0.10 Standard Coalescent

From the simple version of the Wright-Fisher model described in the previous section 0.9.3, where we introduced probabilities concerning different versions of a gene, we move on to the problem of estimating times of when gene variations happen, and from there to derive the *Coalescent Model*, that describes the distribution of the times of the ancestors of different genes sampled from different individuals.

0.10.1 Coalescent of a sample of two different genes

In the same setting of the Wright-Fisher model of section 0.9.3, i.e. a constant population size of N, discrete generation and full mixing of individuals, we want to infer the distribution of the waiting time to the **Most Recent Common Ancestor** (**MRCA**) of two genes sampled in a population of N. Assuming that both genes are sampled at the same time $t = 0$ we will be going backwards in the estimation of the time when they had a common ancestor. Considering discrete generations, the probability that two genes had the MRCA j generations back is given by the probability that they don't have a common ancestor in the previous $j - 1$ generations and they do have a common ancestor in the j_{th} generation: since sampling in different generations is independent of each other, and given the probability $\frac{1}{N}$ that they have a common ancestor in any generation (and therefore $1 - \frac{1}{N}$ that they don't), the time of MRCA is distributed as follows [25] :

$$Pr(T_{MRCA} = j) = \frac{1}{N}(1 - \frac{1}{N})^{j-1} \tag{88}$$

From equation (88) we can see that the time to the common ancestor is geometrically distributed with parameter $\frac{1}{N}$. Equation (88) is derived, under the same assumptions of the Wright-Fisher model, from equation (81): the geometric distribution of (88) comes from the binomial (81) where we focus on the number of "failures" (i.e. the number of generations where there is no coalescent event) until the first "success" (the coalescent event of two samples). Using the properties of the geometric distribution in (88), we can calculate the expected T_{MRCA}

$$E(T_{MRCA}) = \frac{1}{\frac{1}{N}} = N \tag{89}$$

We see, in equation (89), that a bigger population size N means an equally bigger average time to the common ancestor.

0.10.2 Coalescent of a sample of k different genes

We can further expand the expression for two different genes found in equation (88), to the general case of k different genes in a population of N. Under the same conditions explained in the section 0.10.1, if, in starting generation at $t = 0$, out of a population of N, $k \leq N$ individuals have different genes, the probability that in the previous generation they don't have a common ancestor is:

$$\begin{aligned} Pr(T_{MRCA} \neq 1) = (\frac{N-1}{N})(\frac{N-2}{N})...(\frac{N-k+1}{N}) = \\ 1 - \frac{k(k-1)}{2N} + O(\frac{1}{N^2}) \end{aligned} \tag{90}$$

Please note in equation (90), as explained before, that times, in the Coalescent model are counted backwards, therefore $T_{MRCA} = 1$ is one generation back. Since we assume that the population N is significantly larger than k, the term $O(\frac{1}{N^2})$ in (90) can be neglected and therefore the probability of no coalescent events in the previous generation becomes:

$$Pr(T_{MRCA} \neq 1) \approx 1 - \frac{k(k-1)}{2N} \tag{91}$$

And therefore

$$Pr(T_{MRCA} = 1) = 1 - Pr(T_{MRCA} \neq 1) \approx \frac{k(k-1)}{2N} = \binom{k}{2}\frac{1}{N} \tag{92}$$

And, similar to what we derived in equation (88), the probability that two genes out of k different genes in a population of N have a common ancestor j generations back, is given by the probability of no common ancestor for $j-1$ generations, i.e. equation (91) applied $j-1$ times, and then a common ancestor, i.e. equation (92), applied once:

$$Pr(T_{MRCA} = j) = \left(\frac{k(k-1)}{2N}\right)\left(1 - \frac{k(k-1)}{2N}\right)^{j-1} \tag{93}$$

Please note that, in equation (93), for simplicity we have used the $=$ sign instead of $\approx$ as in equations (91) and (92), but it is an approximation nevertheless, valid for $N \gg k$.

0.10.3 The continuous time Coalescent

In the Wright-Fisher model introduced in one of its simplest versions in section 0.9.3 and expanded with equation (93), the time is assumed discrete and indicates the number of generations. Firstly we can notice that, by scaling the time by a factor of N [25] we have that:

$$t_j = \frac{j}{N} \tag{94}$$

Using the time as in (94) allows us to express the results independently from the population size N, and therefore the results will hold for any population, independently of its size, as long as the constraints of the model, explained in sections 0.9.3 and 0.10 hold (importantly that the sample size k is much smaller than the population size N [25]). It can be shown that [25], using the time-scale transformation of equation (94) and the assumption that the population size is much bigger than the number of samples $N \gg k$, the geometric distribution converges to an exponential distribution, and in fact Kingman showed in [26] that as N grows the coalescent process converges to a continuous-time process, having rate shown in the following equation (96)

$$Pr(T_{MRCA} = j) = \left(\frac{k(k-1)}{2N}\right)\left(1 - \frac{k(k-1)}{2N}\right)^{j-1} \xrightarrow{N \gg k} \lambda \exp^{-\lambda j} \tag{95}$$

with the rate of the exponential distribution given by

$$\lambda = \frac{k(k-1)}{2N} \tag{96}$$

Equation (95) gives us the distribution of coalescent times in continuous time. Rewriting (95), and using τ to express the time instead of the discrete j, we have that the density expressing the probability that two lineages out of k coalesce at time τ is given by:

$$Pr(T_{MRCA} = \tau) = exp^{-\frac{k(k-1)\tau}{2N}} = exp^{-\binom{k}{2}\frac{\tau}{N}} \tag{97}$$

The expected value of (97), and therefore the average first coalescent time when we have k different lineages and a population size of N, using the properties of the exponential distribution, is $\frac{1}{\lambda}$, i.e., using equation (96).

$$E(\tau|k, N) = \frac{2N}{k(k-1)} \tag{98}$$

So, using equation (97), if we want to express the density for all the times so that all the k different samples arrive to a unique common ancestor, considering that, as per hypotheses of the generations are not overlapping, there is complete mixing of the population, and that the population size is constant, the probability those times are the result of the coalescent process reducing k lineages into 1 is obtained by multiplying the (independent) probabilities for each coalescence even:

$$f(\tau_0,... \ \tau_k|N) \propto \prod_{i=1}^{k-1} \frac{1}{N} \exp^{-\frac{k_i(k_i-1)\tau_i}{2N}} \tag{99}$$

where, in equation (99), the k_i express the number of different samples at each coalescent event, so for example if we start the analysis with $k = 5$ samples (lineages), at the first coalescent event we will have $k_1 = 5$, then, since two of the lineages will have merged, in the second coalescent event we have $k_2 = 5 - 2 = 3$ lineages, etc, for a number of coalescent events equal to $k - 1$ to arrive to the common ancestor of all.

Bibliography

[1] Annealed Importance Sampling, Radford M. Neal, 1998

[2] Differential Geometric MCMC Methods and Applications, Calderhead B., PhD thesis, University of Glasgow, 2011

[3] The Monte Carlo method, Metropolis N. and Ulam S., Journal of the American Statistical Association, 44(247), 1949

[4] Monte Carlo Methods Lecture Notes, Johansen Adam M. and Evers L., University OfBristol, 2007

[5] An Introduction to MCMC for Machine Learning, Andrieu C. et al, Machine Learning, 50, 543, 2003, Kluwer Academic Publishers

[6] Monte Carlo sampling methods using Markov chains and their Applications, Hastings, W. K. (1970), Biometrika 57, 97109

[7] Equations of state calculations by fast computing machines, Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953), Journal of Chemical Physics, 21, 10871091

[8] Substitution model averaging, Rasmussen D. A., Magnus C., Bouckaert R., website: https://taming-the-beast.org/tutorials/Substitution-model-averaging/

[9] Capturing heterotachy through multi-gamma site models, Bouckaert R., Lockhart P., bioRxiv, Doi: 10.1101/018101 (2015)

[10] Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods, Yang Z., J. Mol Evol (1994) 39:306-314

[11] Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach, Felsenstein J., J Mol Evol (1981) 17:368-376

[12] Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data, Drummond, A. J. , Nicholls, G. K., Rodrigo A. G. and Solomon, W., GENETICS July 1, 2002 vol. 161 no. 3 1307-1320

[13] http://www.who.int/globalchange/environment/en/chapter6.pdf, VVAA

[14] Importance Sampling: Intrinsic Dimension and Computational Cost, Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., Stuart, A. M. Statist. Sci. 32 (2017), no. 3, 405–431. doi:10.1214/17-STS611. https://projecteuclid.org/euclid.ss/1504253124

[15] Monte Carlo Statistical Methods, Robert, C. P., Casella, G. (2004), Springer

[16] Monte Carlo theory, methods and examples, Owen, A. (2013). http://statweb.stanford.edu/owen/mc/

[17] Rethinking The Effective Sample Size, Elvira, V., Martino, L. Robert, C. P. (2018). arXiv: 1809.04129

[18] A note on importance sampling using standardized weights., Kong, A. (1992). University of Chicago, Dept. of Statistics, Tech. Rep 348.

[19] Sequential imputations and Bayesian missing data problems. Kong, A., Liu, J. S. and Wong, W. H. (1994). Journal of the American Statistical Association 9 278-288.

[20] Equations of state calculations by fastcomputing machines, Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953) J. Chem. Phys., 21 (6), 10871092.

[21] Monte Carlo sampling methods using Markov chains and their application, Hastings, W. (1970), Biometrika, 57, 97109.

[22] Statistical Inference in the Wright-Fisher Model Using Allele Frequency Data, Tataru, P., Simonsen M, Bataillon T, Hobolth A., Syst Biol. 2017;66(1):e30-e46. doi:10.1093/sysbio/syw056

[23] Evolution in Mendelian populations., Wright S. (1931). Genetics 16: 97159.

[24] The genetical theory of natural selection., Fisher R.A. (1930). Oxford: Clarendon.

[25] Gene genealogies, variation and evolution: a primer in coalescent theory, Hein, J., Schierup, M. Wiuf, C. (2004), Oxford university press .

[26] The coalescent, Kingman, J. (1982), Stochastic processes and their applications 13(3), 235248

[27] The Curse of Dimensionality for Numerical Integration of Smooth Functions. Hinrichs, A., et al. (2014). Mathematics of Computation. 83. 2853-2863. 10.1016/j.jco.2013.10.007

[28] Importance sampling: Intrinsic dimension and computational cost. Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., Stuart, A., et al. (2017). Statistical Science, 32(3):405431.

[29] Error Bounds and Normalizing Constants for Sequential Monte Carlo in High Dimensions. Beskos, Alexandros , Crisan, Dan , Jasra, Ajay , Whiteley, Nick. (2011)

[30] On the stability of sequential Monte Carlo methods in high dimensions. Beskos, A., Crisan, D., Jasra, A. Ann. Appl. Probab. 24 (2014), no. 4, 1396–1445. doi:10.1214/13-AAP951