

# Importance Sampling for Monte Carlo Evaluation of the Likelihood for Effective Population Size

Eric C. Anderson and Elizabeth A. Thompson  
*University of Washington*

Ellen G. Williamson  
*University of California, Berkeley*

30 June 1999

This Research Supported By  
NSF BIR – 9807747  
NSF BIR – 9256537

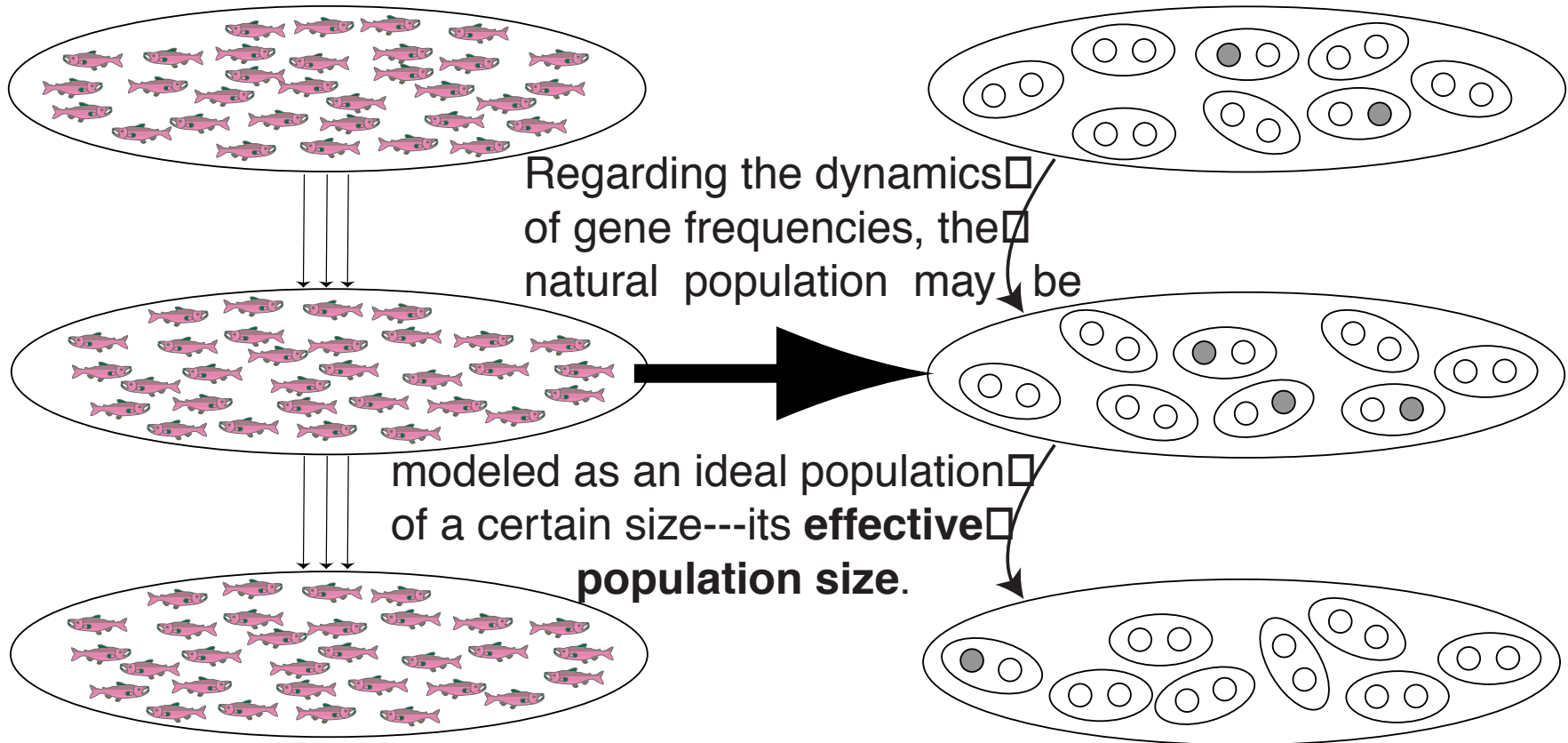
# Overview

- Effective Population Size,  $N_e$
- Estimating  $N_e$  from Allele Frequency Changes
- Why a Likelihood Approach?
- The Likelihood for  $N_e$
- Monte Carlo Evaluation of the Likelihood
  - Importance sampling
  - Approximate Forward-Backward Algorithm
- Analyzing Datasets
  - Simulated
  - Real

# Why A Likelihood Approach?

- MLE has smaller variance than previously developed moment-based estimators
  - Simulation Study with Diallelic Loci (Williamson and Slatkin)
- Extension to More Complex Life Histories
  - Explicit Stochastic Modelling
- Testing Hypotheses about  $N_e$

# Effective Population Size—Translating Natural Populations into W-F Models



**Population Size = 31**

**Effective Population Size = 8**

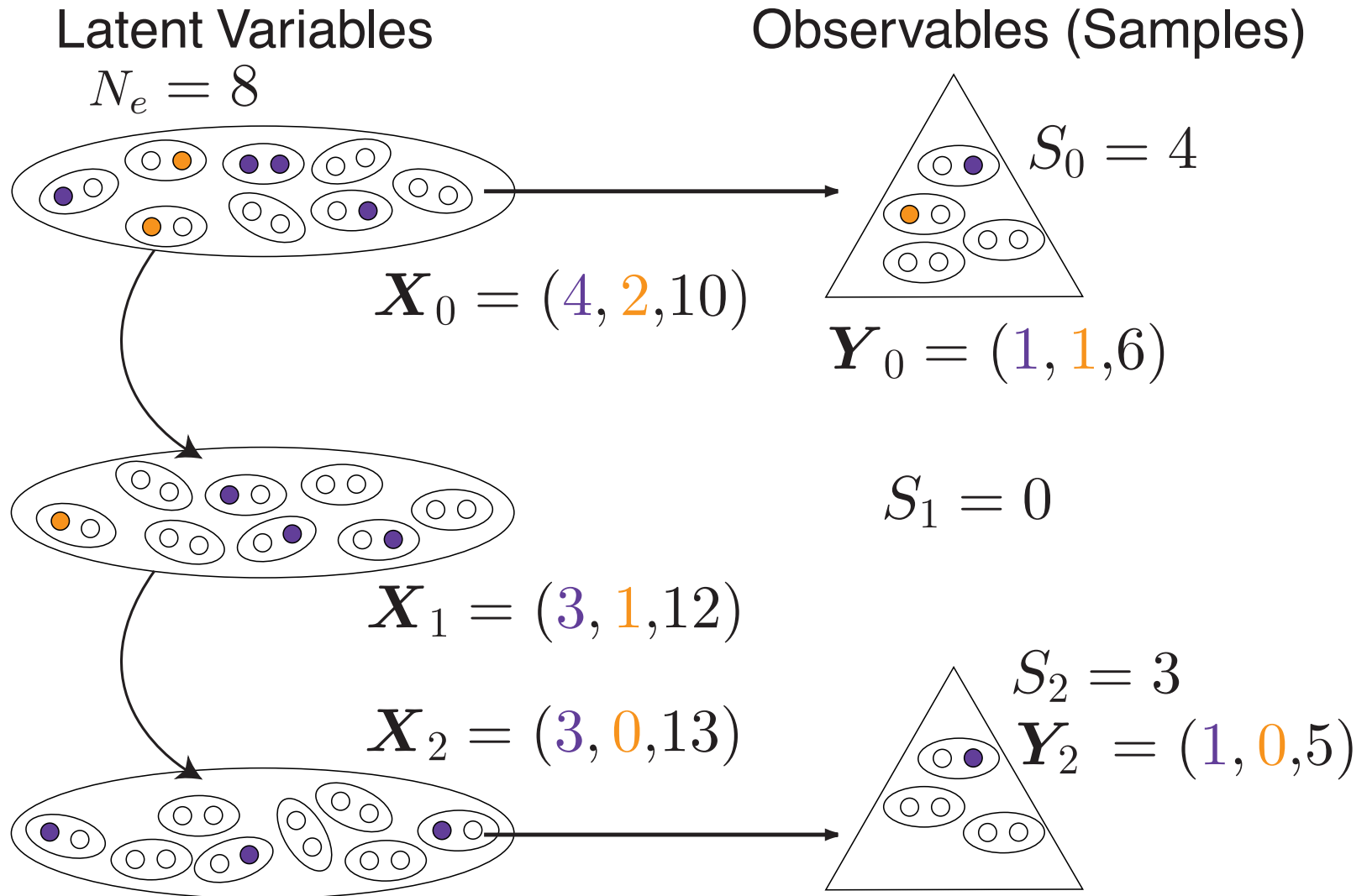
$N_e$

● and ○ are different genetic types (alleles)

# Why *Not* a Likelihood Approach

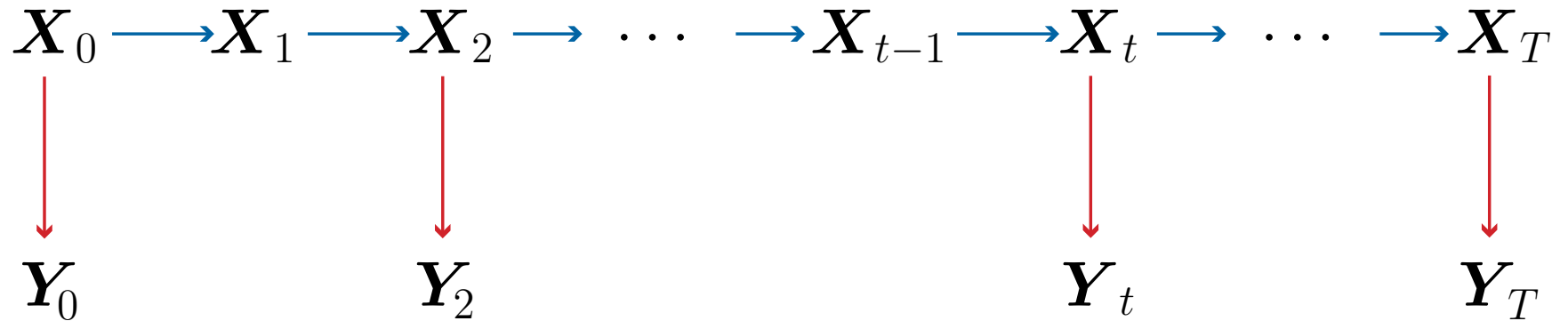
- Exact computation infeasible with more than 2 alleles
- Monte Carlo approximation of the likelihood is not straightforward

# Mathematical Notation



- This depicts a single genetic system called a locus □
- Many loci are available, each giving an independent replicate

## Transition and Sample Probabilities



$$P_{N_e}(\mathbf{X}_t | \mathbf{X}_{t-1}) \sim \text{Mult}_k \left( 2N_e, \frac{\mathbf{X}_{t-1}}{2N_e} \right)$$

$$P_{N_e}(\mathbf{Y}_t | \mathbf{X}_t) \sim \text{Mult}_k \left( 2S_t, \frac{\mathbf{X}_t}{2N_e} \right) \quad \text{if } S_t > 0$$

$k$  is the number of alleles

There are no computational restrictions that  $N_e$  be constant in time

## The Likelihood for $N_e$ given $\mathbf{Y}$

- Note: Integrating out nuisance parameters  $\mathbf{X}_0$
- Assume a (uniform) prior  $\pi(\mathbf{X}_0)$  for them

$$\begin{aligned} L(N_e) &= P_{N_e}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{N_e}(\mathbf{Y}, \mathbf{X}) \\ &= \sum_{\mathbf{x}_0, \dots, \mathbf{x}_T} \pi(\mathbf{X}_0) \left( \prod_{t=1}^T P_{N_e}(\mathbf{X}_t | \mathbf{X}_{t-1}) \right) \left( \prod_{t=0}^T P_{N_e}(\mathbf{Y}_t | \mathbf{X}_t) \right) \end{aligned}$$

The sum is infeasible for more than two alleles

# Monte Carlo Evaluation

$$P_{N_e}(\mathbf{Y}) = \sum_{\mathbf{X}} P_{N_e}(\mathbf{Y}|\mathbf{X})P_{N_e}(\mathbf{X}) = E_{N_e}\left(P_{N_e}(\mathbf{Y}|\mathbf{X})\right)$$

Hence

$$P_{N_e}(\mathbf{Y}) \approx \frac{1}{m} \sum_{i=1}^m P_{N_e}(\mathbf{Y}|\mathbf{X}^{(i)})$$

Where the  $\mathbf{X}^{(i)}$  are drawn from their marginal (unconditional) distribution given  $\pi$  and the probability laws governing their transitions.

# Importance Sampling

$$P_{N_e}(\mathbf{Y}) = \sum_{\mathbf{X}} \frac{P_{N_e}(\mathbf{Y}, \mathbf{X})}{P_{N_e}^*(\mathbf{X})} P_{N_e}^*(\mathbf{X}) = E_{N_e}^* \left( \frac{P_{N_e}(\mathbf{Y}, \mathbf{X})}{P_{N_e}^*(\mathbf{X})} \right)$$

so

$$P_{N_e}(\mathbf{Y}) \approx \tilde{P}_{N_e}(\mathbf{Y}) = \frac{1}{m} \sum_{i=1}^m \frac{P_{N_e}(\mathbf{Y}, \mathbf{X}^{(i)})}{P_{N_e}^*(\mathbf{X}^{(i)})}$$

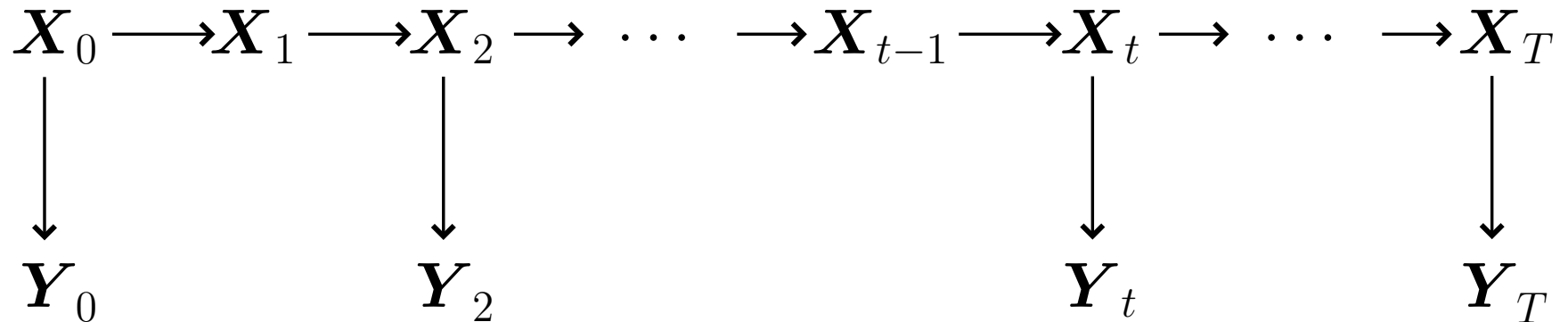
where the  $\mathbf{X}^{(i)}$  are drawn from the distribution  $P_{N_e}^*(\mathbf{X})$ .

- Best  $P_{N_e}^*(\mathbf{X})$  is proportional to  $P_{N_e}(\mathbf{Y}, \mathbf{X})$ .

# Constructing a Suitable $P_{N_e}^*(\mathbf{X})$

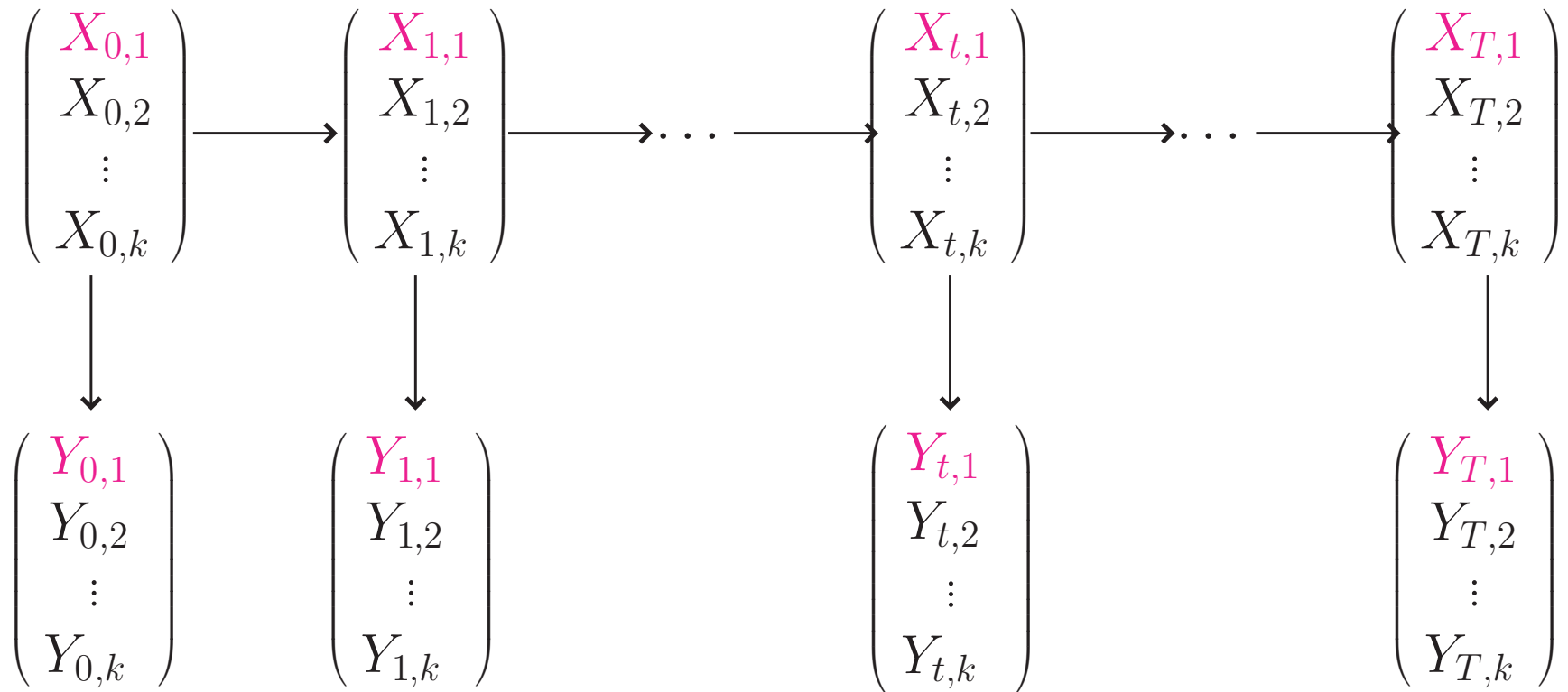
- Must be able to sample from it quickly
- Must be able to compute  $P_{N_e}^*(\mathbf{X})$
- $P_{N_e}^*(\mathbf{X}) > 0$  whenever  $P_{N_e}(\mathbf{Y}, \mathbf{X}) > 0$
- Should be close to proportional to  $P_{N_e}(\mathbf{Y}, \mathbf{X})$

## Baum et al. 1972 Algorithm



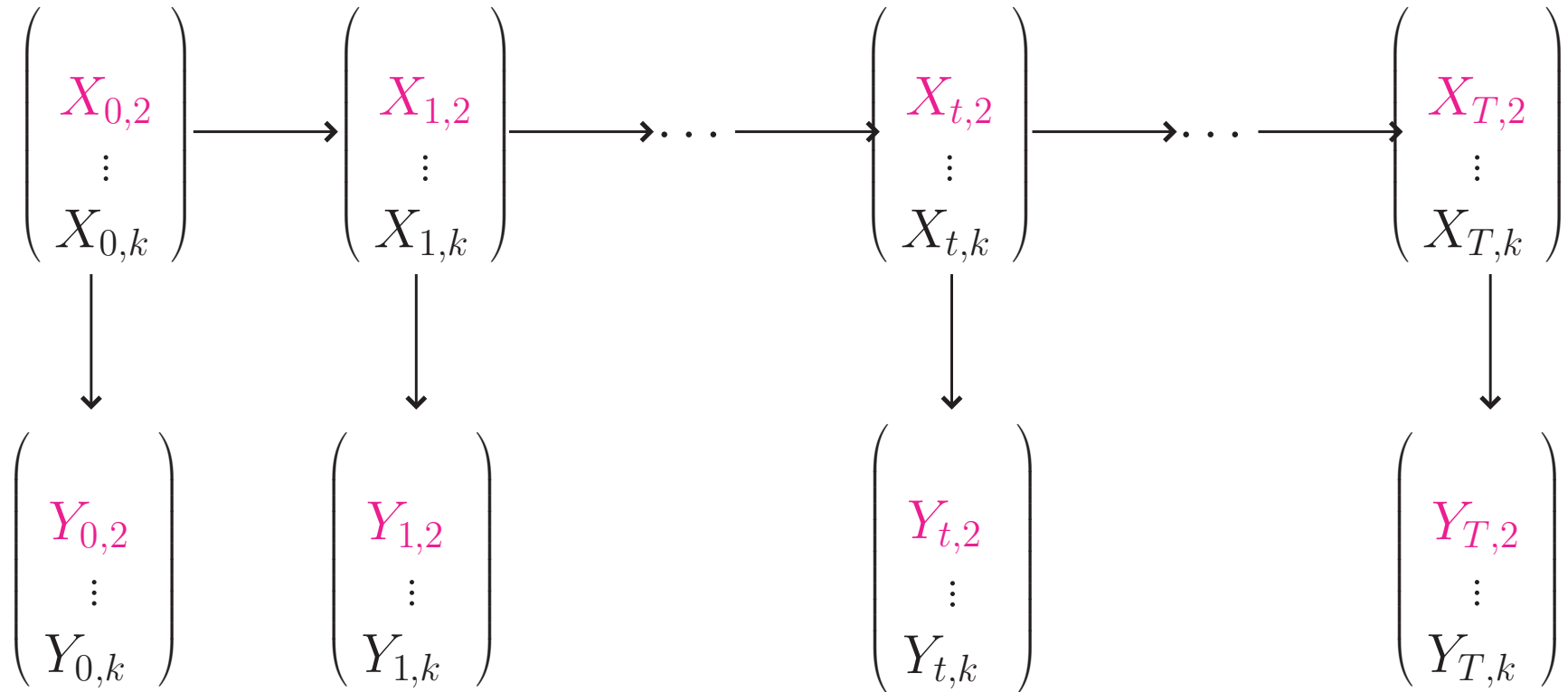
- Forward: Compute and Store  $P_{N_e}(X_t|Y_0, \dots, Y_t)$
- Backward: Realize from  $P_{N_e}(X_t|Y_0, \dots, Y_t, X_{t+1})$
- Delivers a realized value  $\mathbf{X}$  from  $P_{N_e}(\mathbf{X}|\mathbf{Y})$
- Alas, this is also infeasible with more than two alleles

# Sequential Treatment of Alleles



- Realize the components of  $\mathbf{X}$  sequentially over alleles

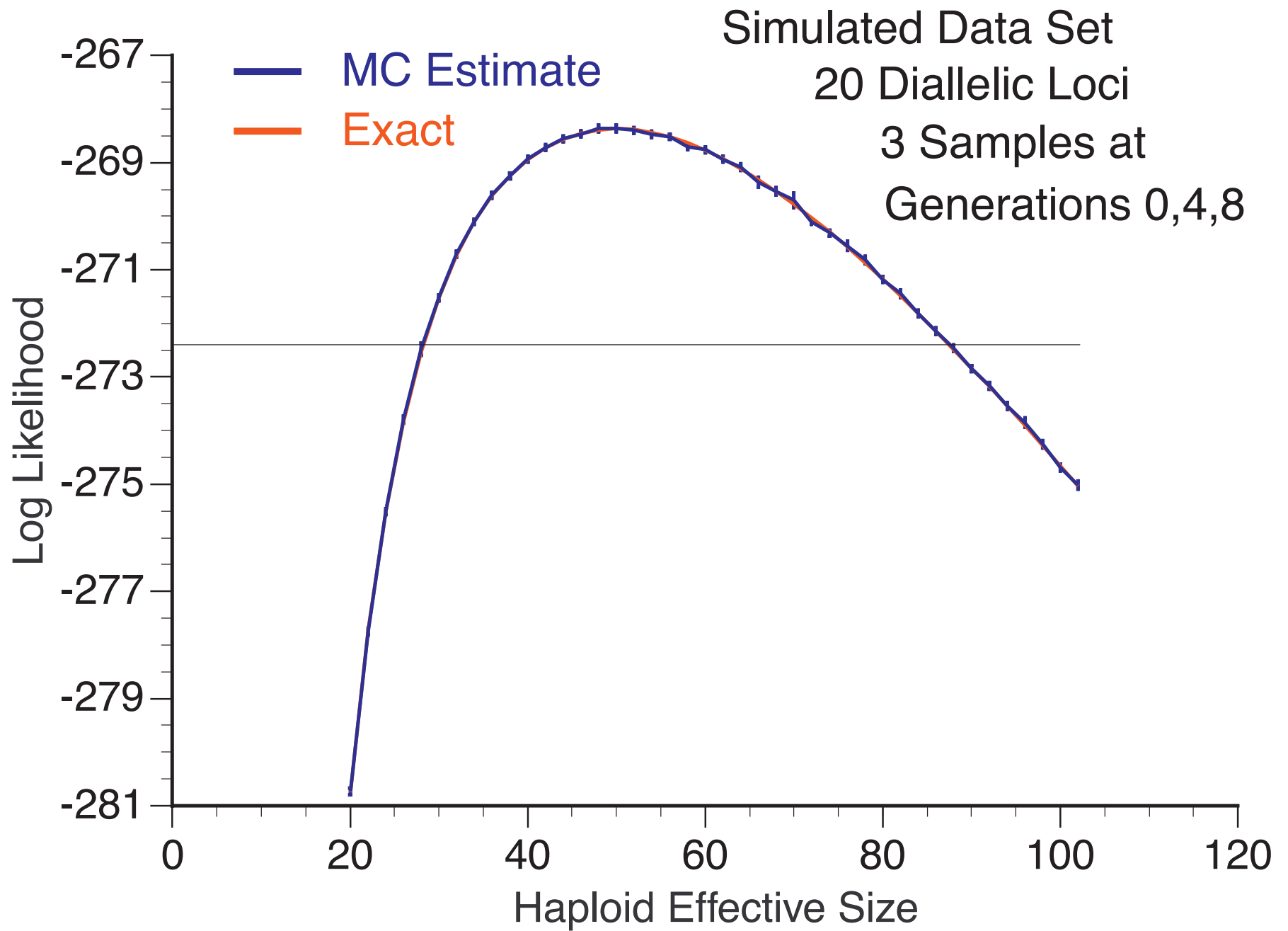
# Sequential Treatment of Alleles

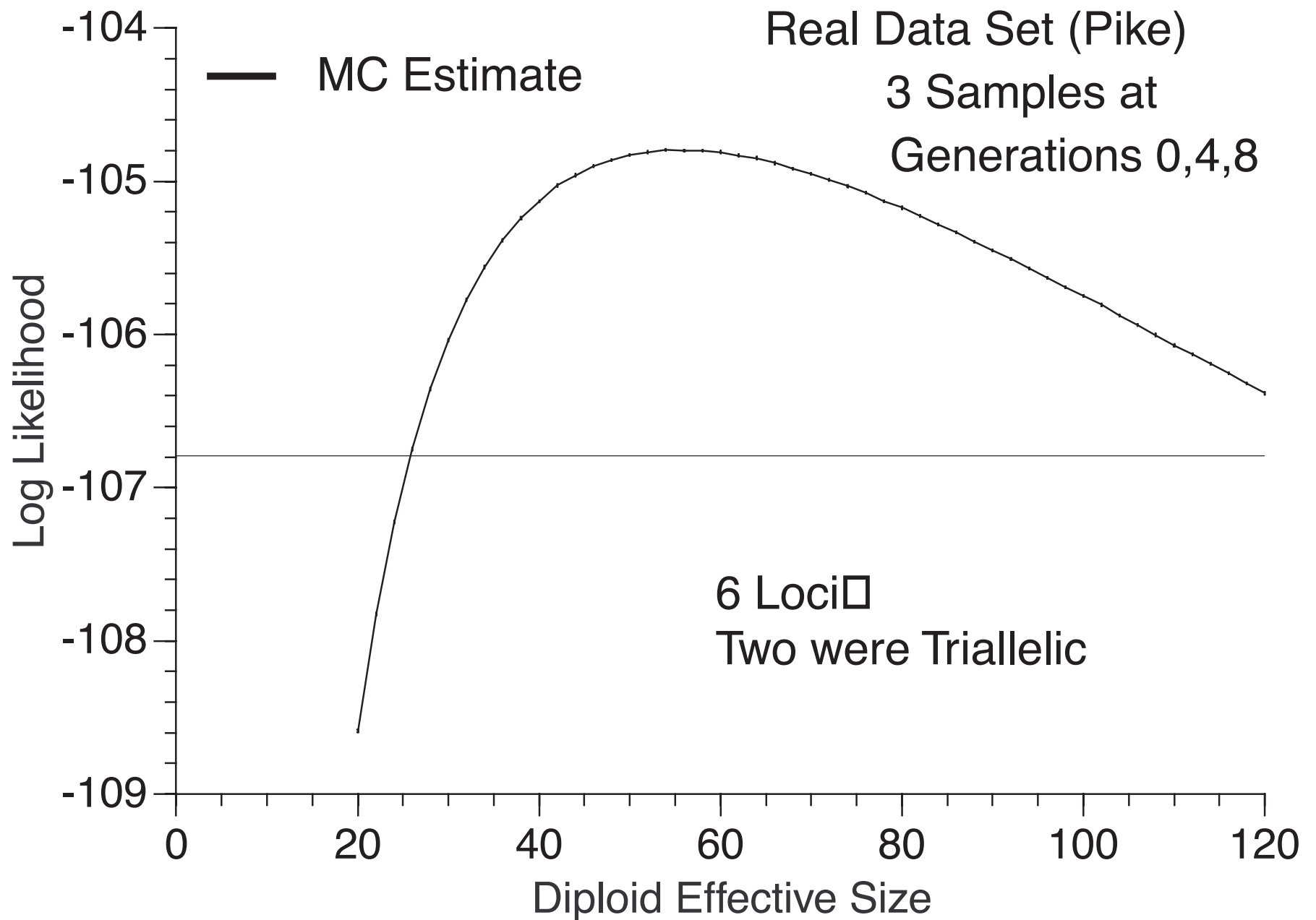


- Adjust  $N_e$ , which is now varying in time:  $(N_{e_1}, \dots, N_{e_T})$
- Adjust  $S_t$ . Then continue until all alleles are realized.

## Normal Approximation for Realizing Alleles

- Even treating alleles sequentially, the sums are onerous.
- So, still sequentially realize alleles, but now:
- Approximate binomial probabilities with normal densities:
  - Sums are transformed into analytically tractable integrals
  - Requires care when transforming back to discrete state space
  - Requires some bookkeeping as well
- Quite fast





Data Source: Miller and Kapuscinski (1997)

# Conclusions

- We require a dsn  $\approx P_{N_e}(\mathbf{X}|\mathbf{Y})$  for importance sampling in Monte Carlo
- State space of  $\mathbf{X}$  too large for exact Baum algorithm
- Employ two approximations to make it computationally feasible
- Performs better than I ever imagined possible