Efficient Likelihood Estimation for Growth Models

Stephanie Keller-Schmidt

Group of Bioinformatics
Group of Parallel Computing and Complex Systems
Departement of Computer Science
University of Leipzig

TBI Winterseminar
Bled/Slovenia February 2010

Outline

Reasons for considering probability models of phylogenetic trees and generate random trees with models :

- Understand speciation and extinction.
- Do predictions that models make about tree shape which can be used to test hypothesis concerning speciation.
- Testing models: how likely is it that model reconstructs a observed tree

Aim: infer how diversity has arisen.
How: fitting stochastic models to tree data.

Databases of Phylogenetic Trees

TreeBASE

- 5212 trees
- leaves are species
- amount of leaves: 4...960
- monotomies and polytomies solved randomly

PANDIT

- 46428 trees
- leaves are proteins
- amount of leaves: 2...5121
- monotomies and polytomies solved randomly

ERM model

Null model of growing trees (simple continous-time branching process).

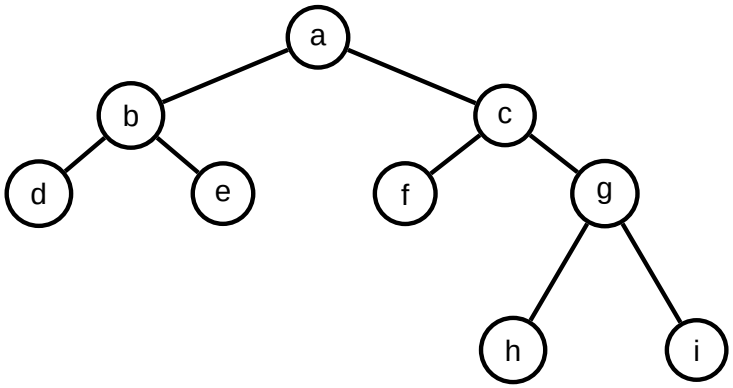
Assumption: Each branch has an equal probability of splitting.

Initialize $t = 0$**:** Generate root with target number of leaves l.

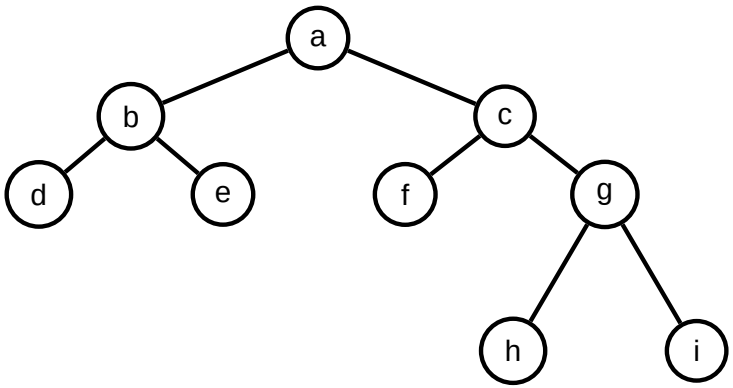
Iterate while $\exists$ leaf l with label $n > 1$:

- Replace leaf l by a cherry.
- Assign new leaves with labels i and $n - i$.
- Probability that the left sister clade contains i taxa is independent of n

Likelihood for ERM model Example

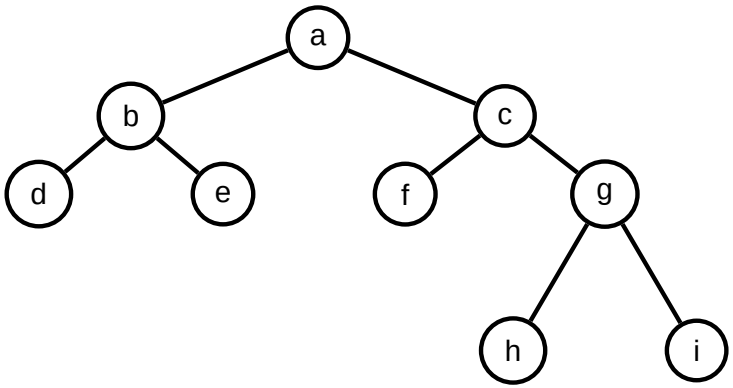


Likelihood for ERM model Example



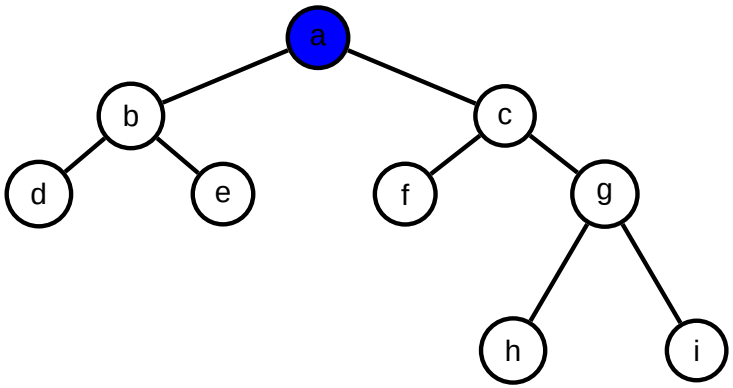
$$L_{\mathrm{ERM}}(T) = \prod_{x \in I(T)} p_{\mathrm{A}}(s(\mathrm{left}(x))|s(x))$$

Likelihood for ERM model Example



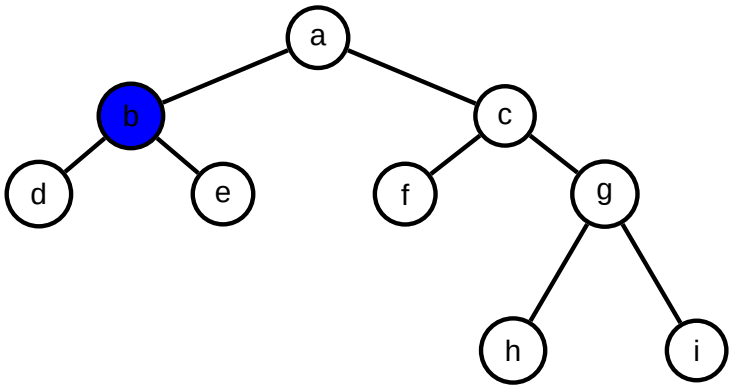
$$L_{ERM}(T) =$$

Likelihood for ERM model Example



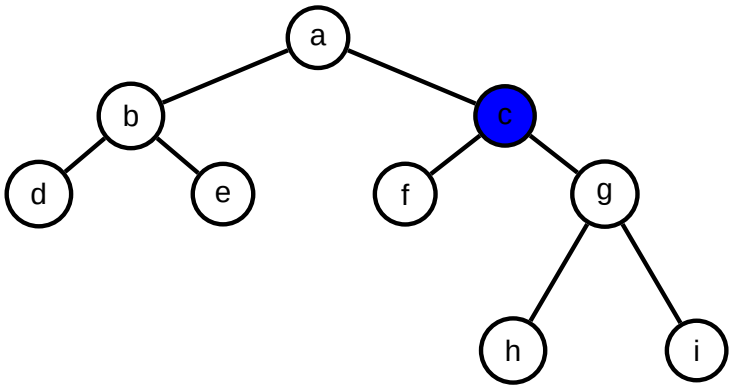
$$L_{ERM}(T) = p_a(2|5)\cdot$$

Likelihood for ERM model Example



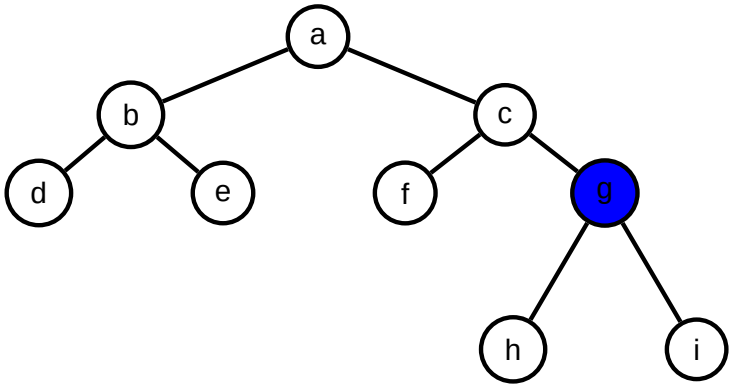
$$L_{ERM}(T) = p_a(2|5) \cdot p_b(1|2) \cdot$$

Likelihood for ERM model Example



$$L_{ERM}(T) = p_a(2|5) \cdot p_b(1|2) \cdot p_c(1|3) \cdot$$

Likelihood for ERM model Example



$$L_{ERM}(T) = p_a(2|5) \cdot p_b(1|2) \cdot p_c(1|3) \cdot p_g(1|2)$$

Age model

Idea: The longer species i has not been involved in speciation, the less likely it is to do so now.

Initialize: Set time $t = 0$, generate root node.

Iterate:

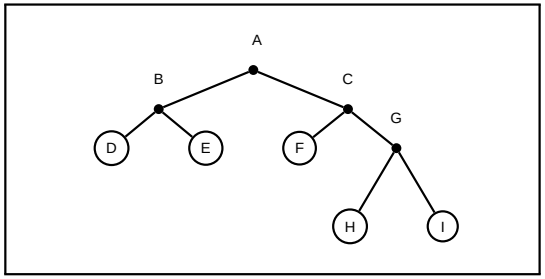
- Increment time t.
- From the set of leaves, choose leaf l with probability

$$p_i \propto (t - t_l)^{-1}$$

- Replace l by a cherry.

t = number of leaves = current time; t_l creation time of leaf l

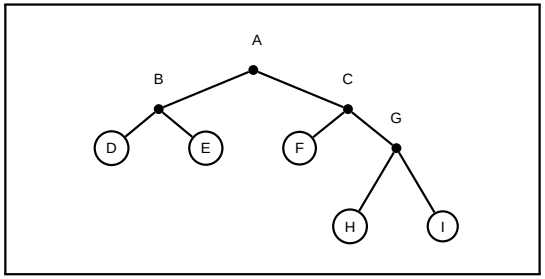
Age model - Example



$t = 0$

$L_{t=0} = \{A_{age=0}\}$

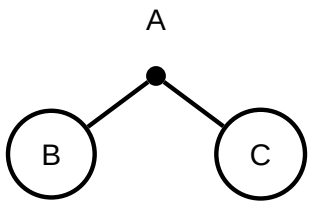
Age model - Example



$t = 1$

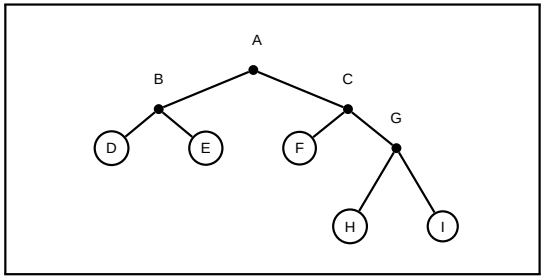
old : $L_{t=0} = \{A_{age=0}\}$

new: $L_{t=1} = \{B_{age=0}, C_{age=0}\}$



$$P_A = \frac{1}{(1-0)} = 1$$

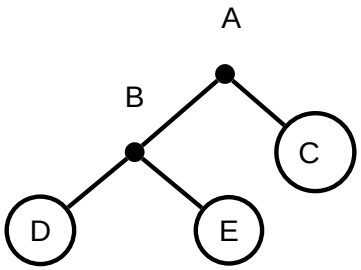
Age model - Example



$t = 2$

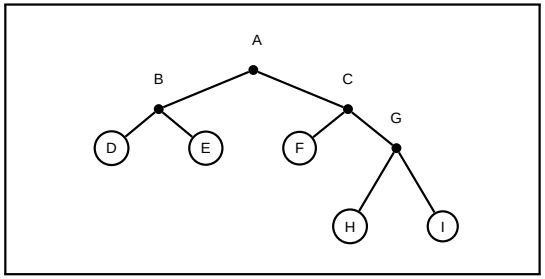
old : $L_{t=1} = \left\{B_{age=0}, C_{age=0}\right\}$

new: $L_{t=2} = \left\{C_{age=1}, D_{age=0}, E_{age=0}\right\}$



$$P_B = \frac{1}{(2-0)} = \frac{1}{2}$$

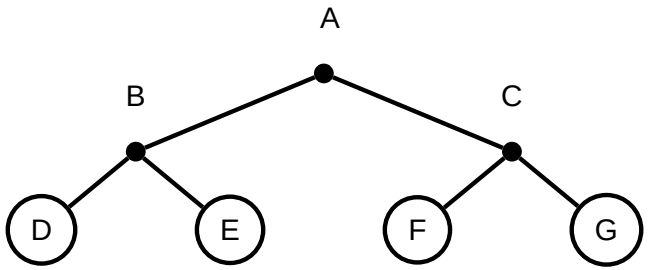
Age model - Example



$t = 3$

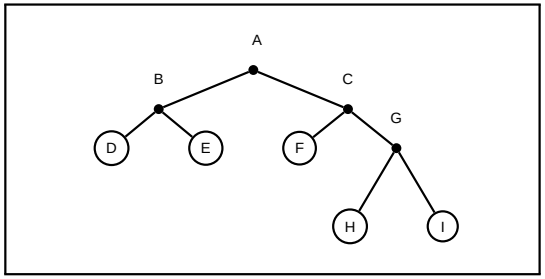
old : $L_{t=2} = \left\{C_{age=1}, D_{age=0}, E_{age=0}\right\}$

new: $L_{t=3} = \left\{D_{age=1}, E_{age=1}, F_{age=0}, G_{age=0}\right\}$



$$P_C = \frac{1}{(3-1)} = \frac{1}{2}$$

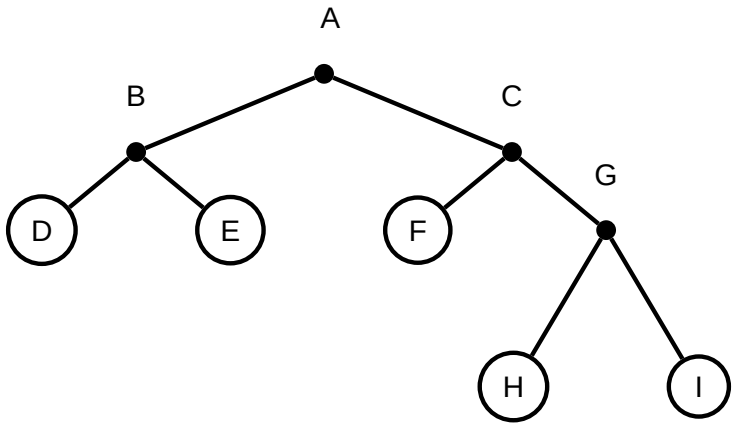
Age model - Example



$t = 4$

old : $L_{t=3} = \left\{D_{age=1}, E_{age=1}, F_{age=0}, G_{age=0}\right\}$

new: $L_{t=4} = \left\{D_{age=2}, E_{age=2}, H_{age=0}, I_{age=0}\right\}$



$$P_G = \frac{1}{(4-0)} = \frac{1}{4}$$

Likelihood - Exact Calculation

- For ERM model

$$L_{\mathrm{ERM}}(T) = \prod_{x \in I(T)} p_{\mathrm{A}}(s(\mathrm{left}(x)) | s(x))$$

- For AGE model
  - Calculate $P_{AGE}(T)$ exactly by adding up probabilities of all sequences of branchings for T

$$L_{\mathrm{AGE}}(T) = \sum_{s \in S_c(t)} p(s,T)$$

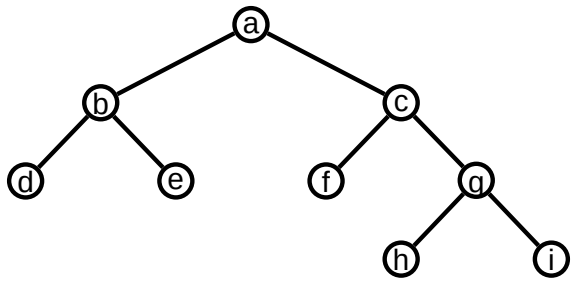
with

$$p(s,T) = \prod_{i=2}^{n-1} \frac{(s(i) - s(m(i))^{-1}}{\sum_{j \in B(s,s(i))} (s(i) - s(m(j))^{-1}}$$

and

$$B(s,t) = \{j \in I \setminus \{1\} \mid s(m(j)) < t < s(j)\} \cup \{j \in A \setminus I \mid s(m(j)) < t\}$$

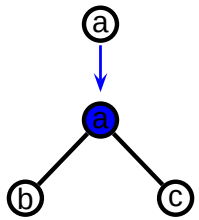
Likelihood for AGE model Example



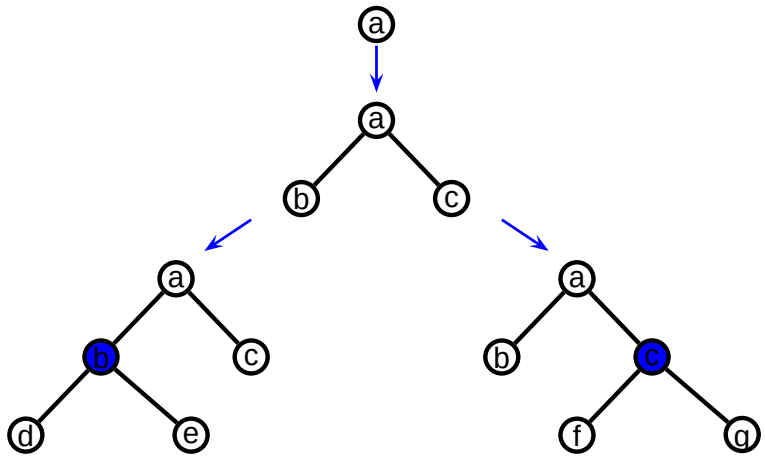
$$L_{\mathrm{AGE}}(T) = \sum_{s \in S_c(t)} p(s, T)$$

Likelihood for AGE model Example

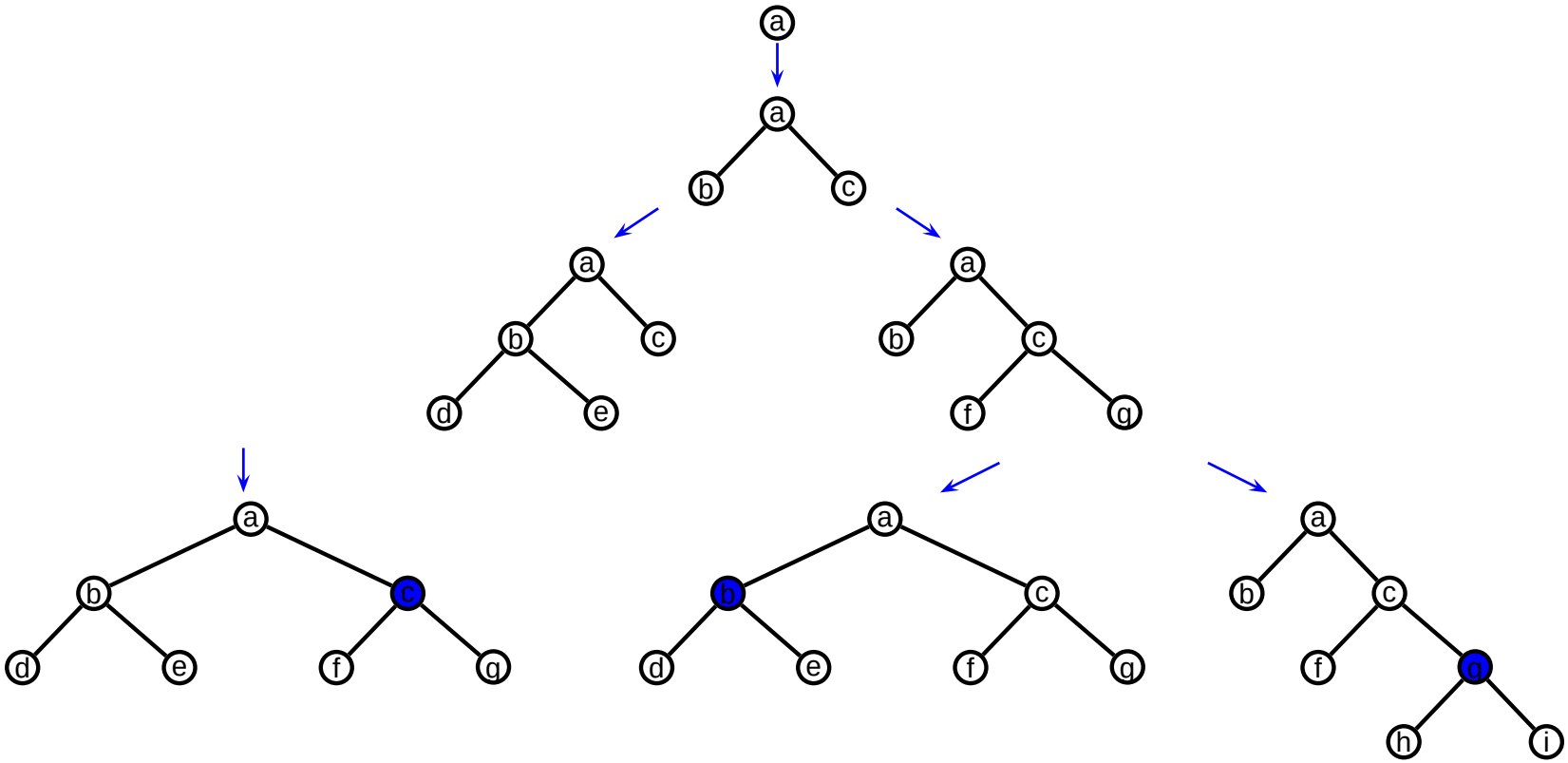
Likelihood for AGE model Example



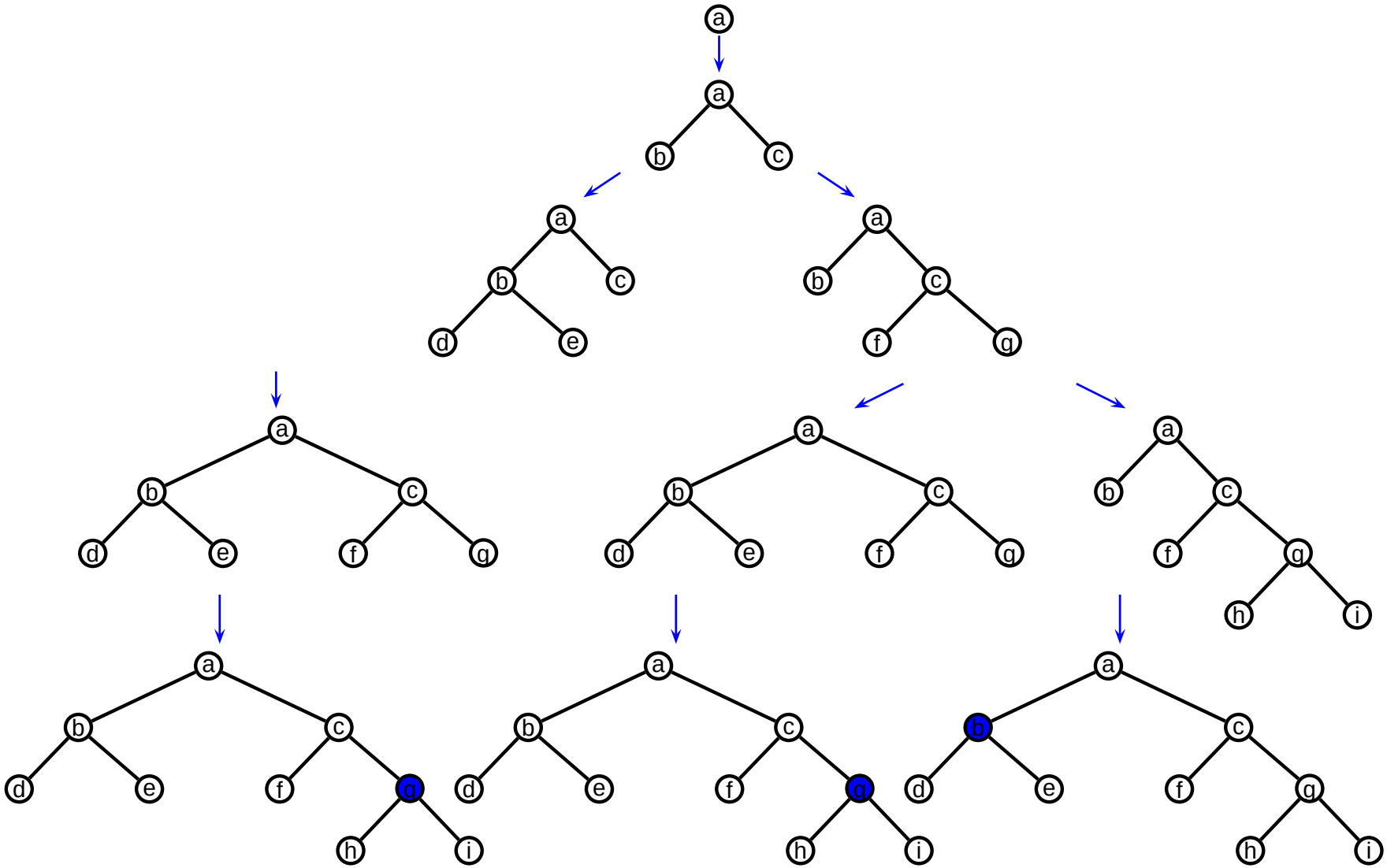
Likelihood for AGE model Example



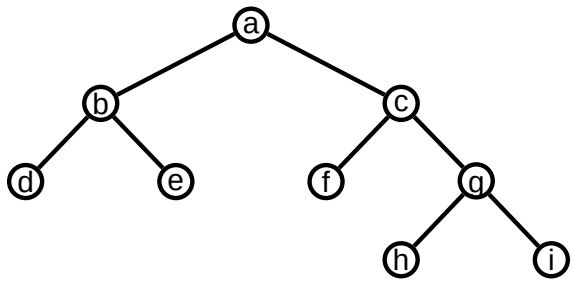
Likelihood for AGE model Example



Likelihood for AGE model Example



Likelihood for AGE model Example



$$L_{\mathrm{AGE}}(T) = \sum_{s \in S_c(t)} p(s,T)$$

$$L_{\mathrm{AGE}}(T) = p((b,c,g),T) + p((c,b,g),T) + p((c,g,b),T)$$

Likelihood - Estimation for growth models

Naive ways of sampling:

- Enough calculation capacity

Or

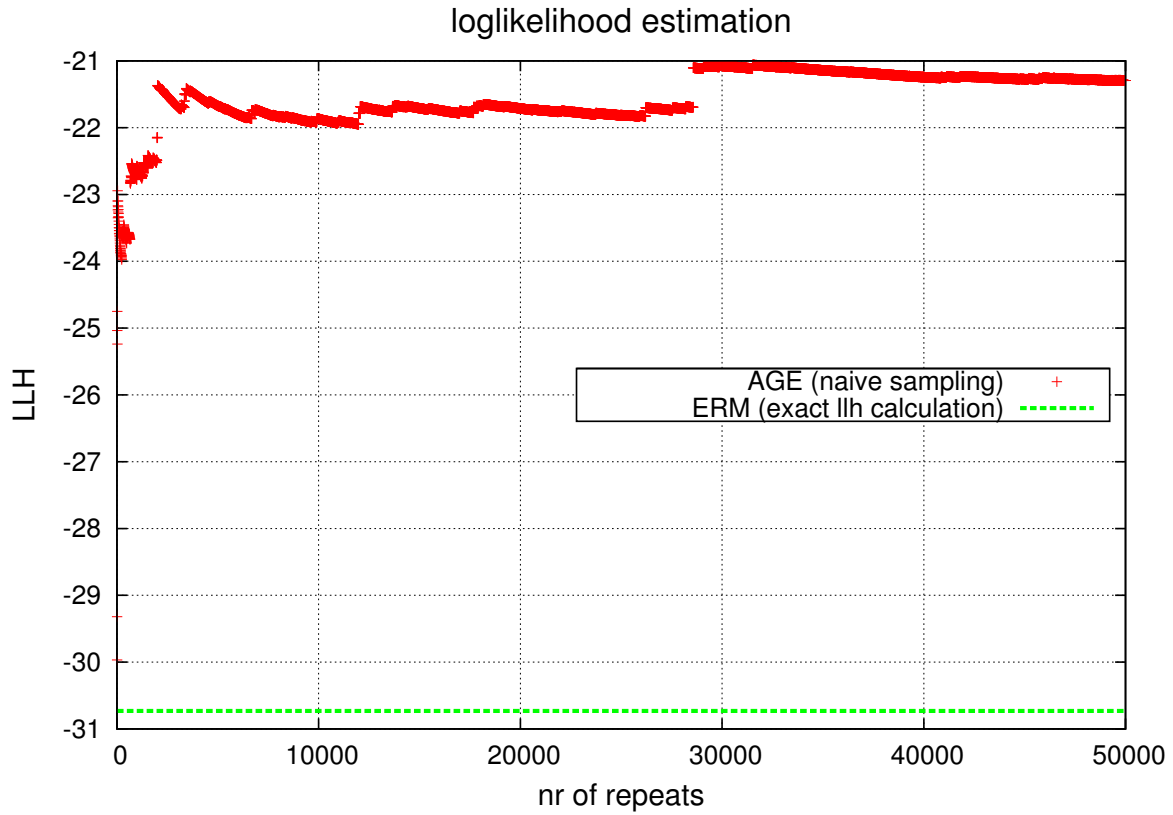
- B is set of all branching sequences leading to “target tree”
- $C \subseteq B$ is sample of B with $|C| << |B|$
- Each possible path has same probability

$$L = \frac{|B|}{|C|} * \sum_{\vartheta \in C} p(\vartheta)$$

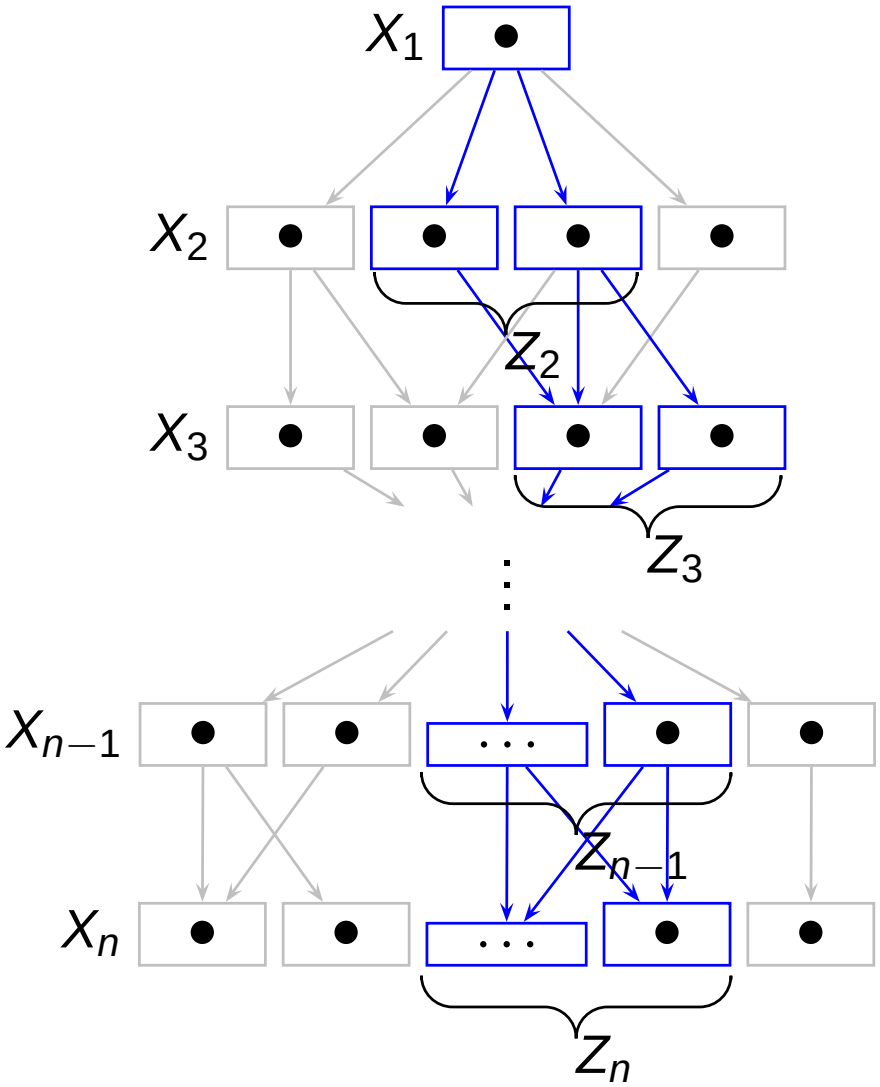
Likelihood - Estimation for growth models

What if $|Z|$ is too large?

Naive Approach: Sample each trajectory with equal probability.
Problem: # trajectories $\uparrow$ and # samples $\downarrow$ | small $Z_n \rightsquigarrow L(\Theta \in Z_n) \downarrow$



Likelihood - Estimation for growth models



$\Rightarrow$ q-dynamics restricted to $Z_1, \ldots Z_n$

$$q_i(y|x) = \frac{p_i(y|x)}{s(x)},$$

$i \in \{1, \ldots n-1\}$, $x \in Z_i$, $y \in Z_{i+1}$

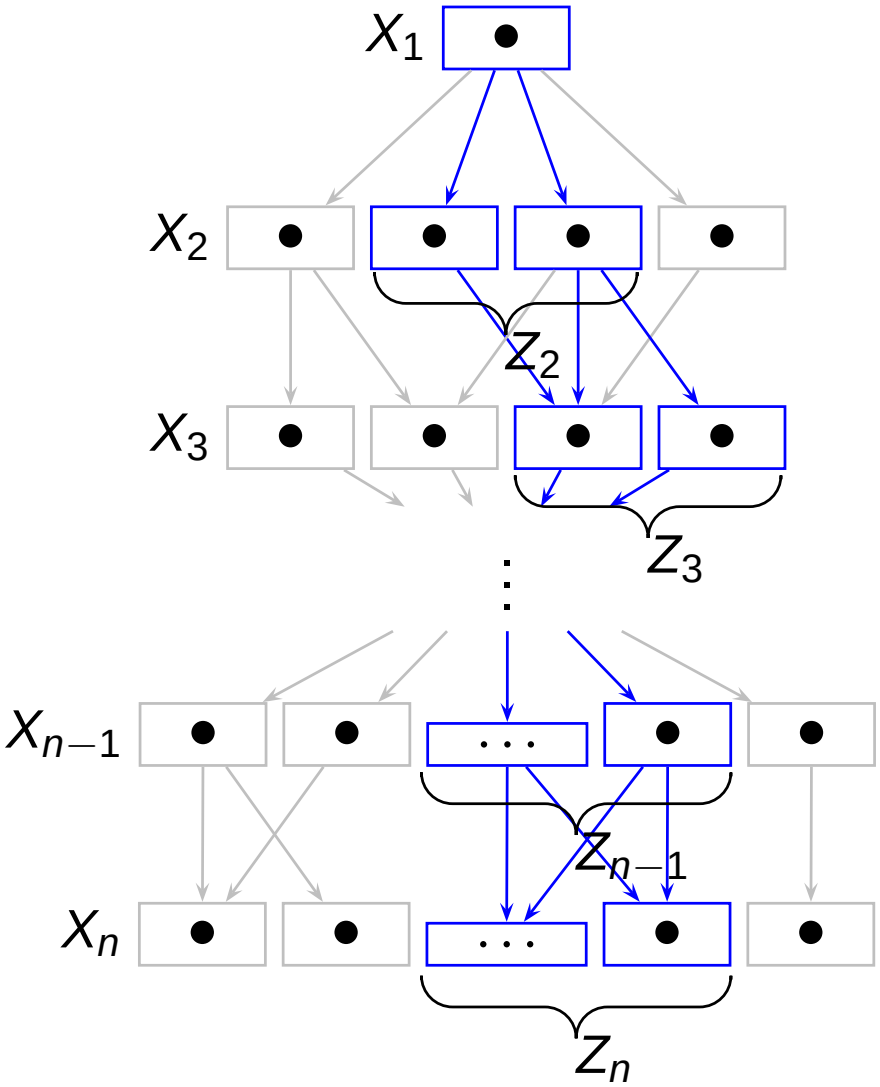
$\Rightarrow$ normalization

$$s(x) = \sum_{y \in Z_i + 1} p(y|x) \, .$$

$\Rightarrow$ probability with which system produces $\Theta \in Z$

$$S(\Theta) = \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i)$$

Likelihood - Estimation for growth models



$\Rightarrow$ q-dynamics restricted to $Z_1, \ldots Z_n$

$$q_i(y|x) = \frac{p_i(y|x)}{s(x)},$$

$$i \in \{1, \ldots n-1\},\ x \in Z_i,\ y \in Z_{i+1}$$

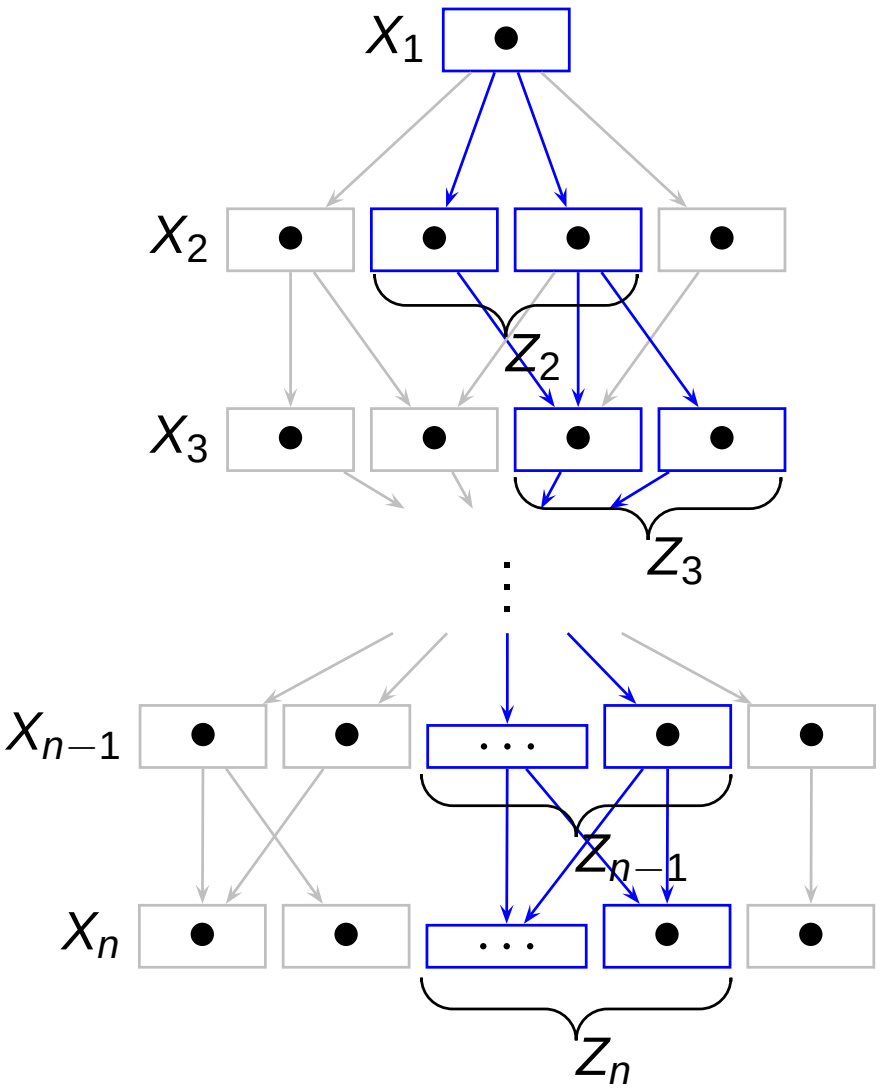
$\Rightarrow$ normalization

$$s(x) = \sum_{y \in Z_i+1} p(y|x)\ .$$

$\Rightarrow$ probability with which system produces $\Theta \in Z$

$$S(\Theta) = \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i)$$

Likelihood - Estimation for growth models



$\Rightarrow$ q-dynamics restricted to $Z_1, \ldots Z_n$

$$q_i(y|x) = \frac{p_i(y|x)}{s(x)},$$

$$i \in \{1, \ldots n-1\},\ x \in Z_i,\ y \in Z_{i+1}$$

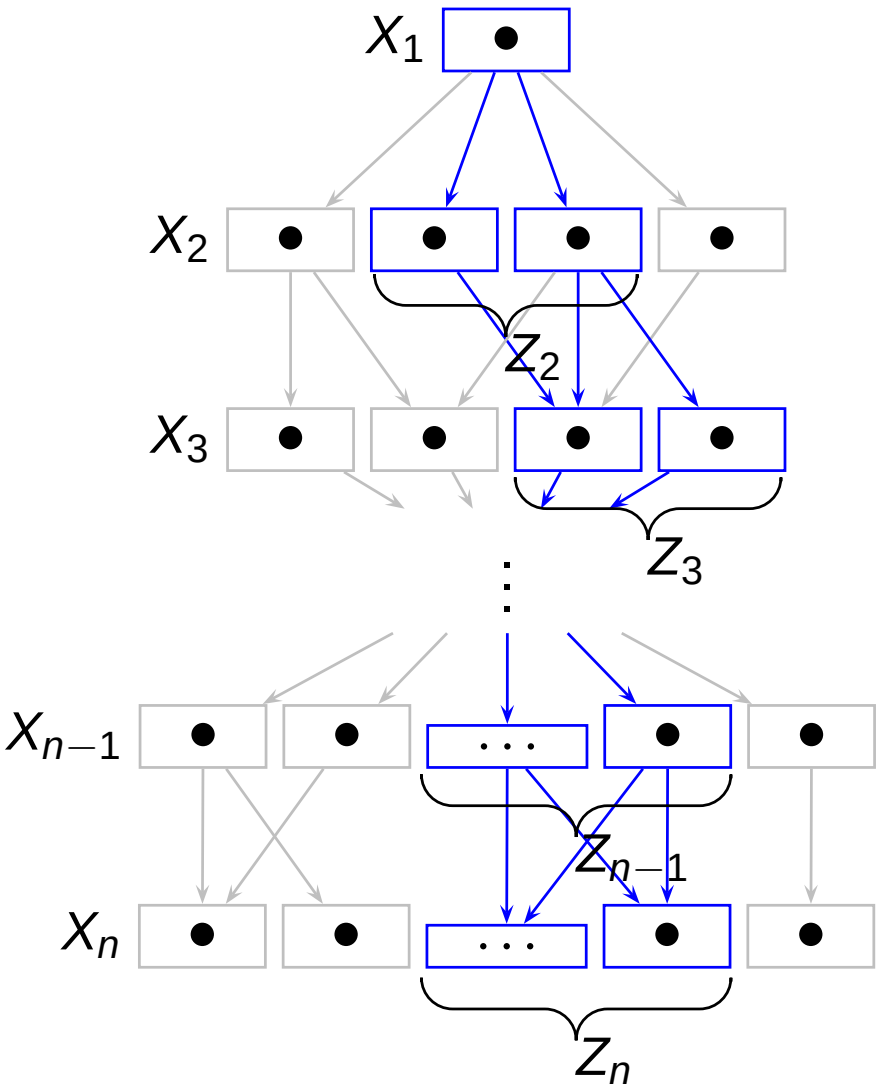
$\Rightarrow$ normalization

$$s(x) = \sum_{y \in Z_i + 1} p(y|x)\ .$$

$\Rightarrow$ probability with which system produces $\Theta \in Z$

$$S(\Theta) = \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i)$$

Likelihood - Estimation for growth models



$\Rightarrow$ q-dynamics restricted to $Z_1, \ldots Z_n$

$$q_i(y|x) = \frac{p_i(y|x)}{s(x)},$$

$$i \in \{1, \ldots n-1\},\ x \in Z_i,\ y \in Z_{i+1}$$

$\Rightarrow$ normalization

$$s(x) = \sum_{y \in Z_i + 1} p(y|x)\ .$$

$\Rightarrow$ probability with which system produces $\Theta \in Z$

$$S(\Theta) = \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i)$$

Likelihood - Estimation for growth models

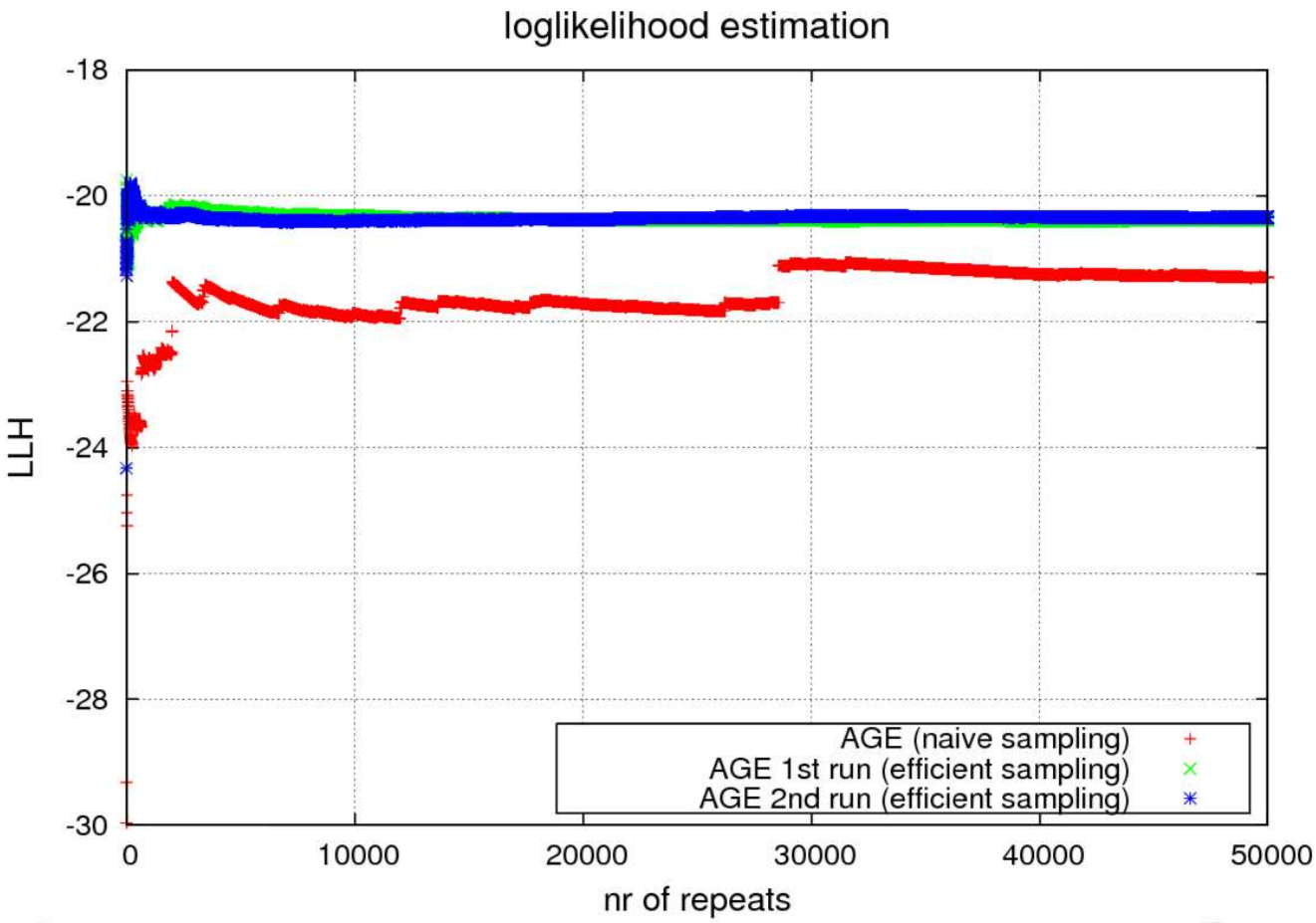
Assign “output" A for each trajectory $\Theta \in Z$

$$A(\Theta) = \prod_{i=1}^{n-1} s(\Theta_i) \; .$$

Expectation value of A over trajectories under q-dynamics = probability L that p-dynamics ends up in the target set Z_n.

$$\Rightarrow \langle A \rangle = L$$

Likelihood - Estimation for growth models



Summary

- Sample loglikelihood of growth models using an importance sampling method.
- Applicable if for each $i \in \{1, \ldots, n-1\}$ and all states $x \in X_i$
  1. it can be decided efficiently (fast) if $x \in Z_i$ or not.
  2. the normalization $s(x)$ can be computed efficiently.
- Requirements are fullfilled by the models of tree growth
  $\Rightarrow$ Use the most probable branching sequences only.

Thanks to Konstantin

and

Expectation value of A over trajectories under q-dynamics = probability L that the p-dynamics ends up in the target set Z_n, as shown by the following sequence of term replacments.

$$\langle A \rangle = \sum_{\Theta \in Z} S(\Theta) A(\Theta) \tag{1}$$

$$= \sum_{\Theta \in Z} \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i) \prod_{j=1}^{n-1} s(\Theta_j) \tag{2}$$

$$= \sum_{\Theta \in Z} \prod_{i=1}^{n-1} q_i(\Theta_{i+1}|\Theta_i) s(\Theta_i) \tag{3}$$

$$= \sum_{\Theta \in Z} \prod_{i=1}^{n-1} p_i(\Theta_{i+1}|\Theta_i) \tag{4}$$

$$= \sum_{\Theta \in Z} R(\Theta) \tag{5}$$

$$= \sum_{\Theta \in X} R(\Theta) \tag{6}$$

$$= L \tag{7}$$