Markovian Models of Genetic Inheritance – Lecs 3,4

Correlation Decay and Phylogenetic Reconsruction

Elchanan Mossel,

U.C. Berkeley

mossel@stat.berkeley.edu,
http://www.cs.berkeley.edu/~mossel/

Can we do better? Information Decay

- Conclusion of last lecture: Impossible to reconstruct if $k \leq 0.5 \log n$ and possible if $k \geq n^{\alpha}$. What is the truth?
- This lecture we will consider this problem and it's relation to correlation decay.

- Def: Consider a Phylogenetic model $T=(V,E,P,r,L)$ rooted at $r \in V$ and with set leaves L.
- For $a \in \Sigma$, let $P^a = P \mid \sigma(r) = a$
- Let $Q^a = P^a$ on the algebra generated by $\sigma(L)$.
- Let $\eta(T,r) = \min_Q \max_a |Q^a - Q|_{TV}$
- Informally measures information from leaves on root.
- Related to the "reconstruction problem".

Information Decay and Reconstruction

- Def: Let $\eta(T,r) = \min_Q \max_a |Q^a - Q|$

- Thm: Consider the Phylogenetic Reconstruction problem
- for balanced binary trees where
- all edges have identical Markov processes and
- assuming a uniform prior over trees.
- Then the probability of correct reconstruction of trees of depth $r+s$ from sequences of length k is at most

  $2^s\, k\, \eta(T,r) + 1 / N_s$
- where N_s is the number of balanced binary trees of s levels on 2^s labeled leaves.

Information Decay and Reconstruction

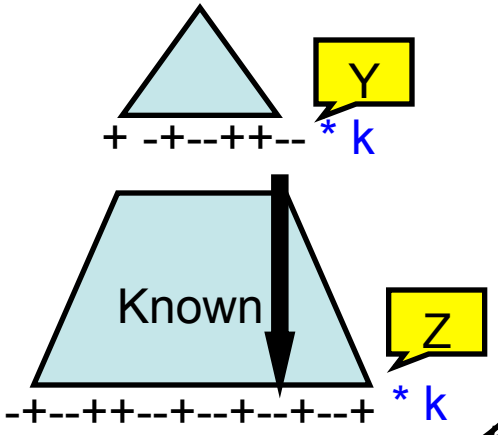
- Thm: The probability of correct reconstruction of trees of depth $r+s$ from sequences of length k is at most $2^s k \eta(T,r) + 1 / N_s$

- Cor: To reconstruct with probability 0.9 need
- $k \geq 2^{-5} / \eta(T,r-4)$
- Later we'll see that for some models $\eta(T,r) \leq 0.5^{cr}$ for some $c > 0$. For these models we need:
- $k \geq 2^{cr-5} = 2^{-5} n^c$
- For some models polynomial sequence length is needed.
- Related papers: M-03,M-04,M-Roch-Sly-11.

Proof of Lower Bounds

- Thm: The probability of correct reconstruction of trees of depth $r+s$ from sequences of length k is at most $2^s k \eta(T,r) + 1 / N_s$
- Pf: Assume: topology of top s levels is chosen uniformly at random and topology of bottom r levels is given.
- From the assumptions it follows that there exists a measure Q on the leaves such that $Q_T = (1-\tau) Q + \tau R_T$
- $\tau \leq 2^s k \eta$ and Q is independent of the tree. Now:
- P[Correct recon] =
- $(1-\tau)$ E Q[Correct recon] +

  τ E Q_T[Correct recon]
- $\leq (1-\tau)/N_s + \tau \leq 2^s k \eta(T,r) + 1 / N_s$
- Can take Q = product measure.

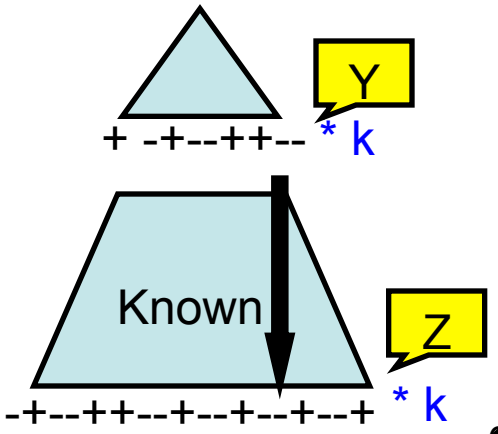


Proof of Lower Bound: Details

- Details completed:
- Let Q' be such that for the r level tree:
- $Q^a = (1-\eta)\, Q' + \eta\, R^a$ (Q' doesn't depend on a).
- Then for the $r+s$ level tree we may write:
- $Q'_T = (1-2^s\,\eta)\, Q'' + 2^s\,\eta\, R'_T$
- Q'' is just Q' to the power 2^s
- Similarly: $Q_T = Q'_T \times \ldots \times Q'_T =$

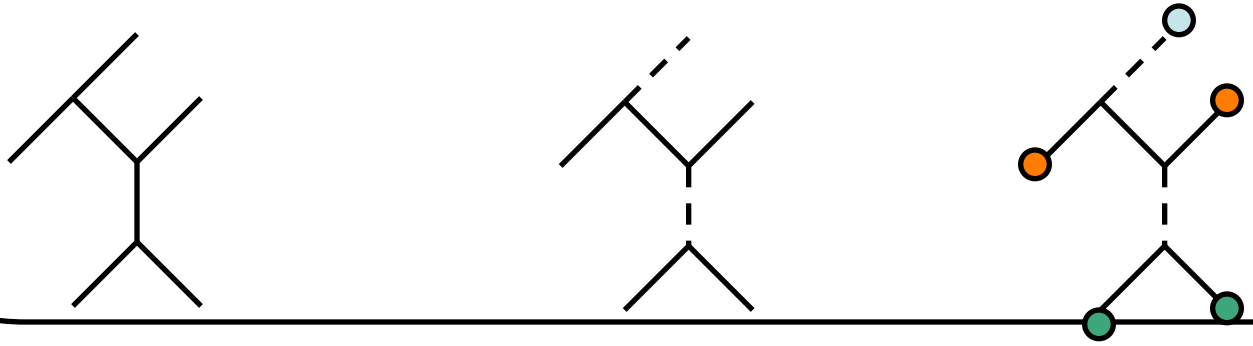
  $(1- k\,2^s\,\eta)\, Q + k\,2^s\,\eta\, R_T.$

  where Q is a power of Q'



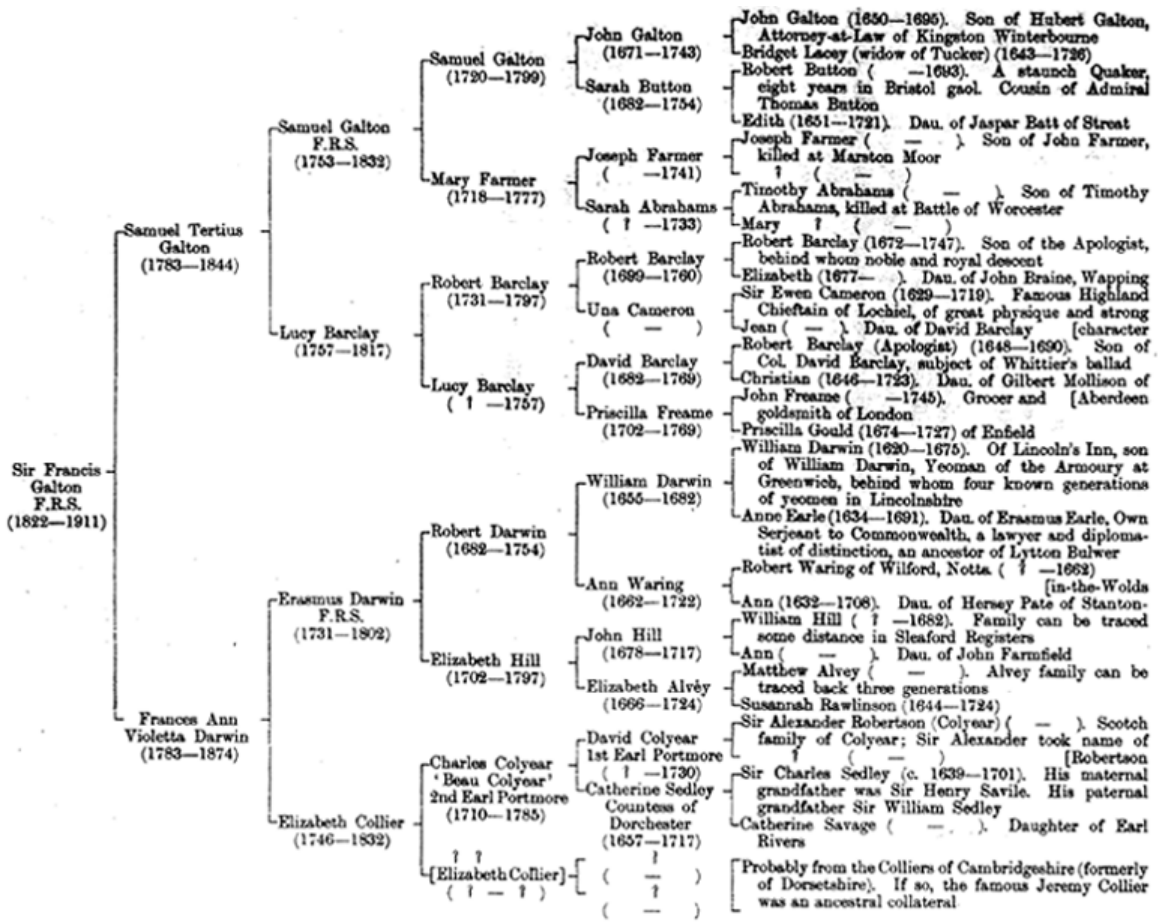
The "random cluster" model

- Infinite set A of colors.
  - "real life" – large $|A|$; e.g. gene order.
- Defined on an un-rooted tree $T=(V,E)$.
- Edge e has (non-mutation) probability $\theta(e)$.
- Character: Perform percolation – edge e open with probability $\theta(e)$.
- All the vertices v in the same open-cluster have the same color σ_v. Different clusters get different colors. This is the "random cluster" model (both for (P,V,E) and $(P^{\otimes k}, V, E)$



Galton-Watson

Galton was interested in genetic explanations of why he was so brilliant.



Correlation Decay for "random cluster" models

- Claim: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then the probability that
- there exists a leaf u which is a descended of v, with $\sigma(v) = \sigma(u)$ is at most $3\,(1 - 2\,\varepsilon)^{d(v,L)}$
- where $d(v,L)$ is the tree distance between v and the leaf closest to v.
- Proof:
- Each leaf u has a probability at most $(\frac{1}{2} - \varepsilon\,)^{d(v,u)}$ of having the same color as v.
- The result follows by a union bound.

"random cluster" model reconstruction

- Claim: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then the probability that
- there exists a leaf u which is a descended of v, with $\sigma(v) = \sigma(u)$ is at most $3\,(1 - 2\,\varepsilon)^{d(v,L)}$
- where d(v,L) is the tree distance between v and the leaf closest to v.
- Cor: Suppose $\theta(e) = \frac{1}{2} - \epsilon$ for all e and that T is a balanced binary tree of l levels rooted at r then:
- $\eta(T,l) \leq 3\,(1 - 2\,\varepsilon)^{l}$
- Pf Sketch: Let Q be the RC measure conditioned on having no path from the root to any of the leaves.
- Cor: The Phylogenetic reconstruction problem requires $k \geq 2^{-7}\,(1 - 2\,\varepsilon)^{-l} = 2\text{^}\{-7\}\ n^{\log\_2(1/(1-2\,\varepsilon))}$ samples in order to reconstruct the tree with probability at least 0.9.

"random cluster" model summary

- Summary: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then:
- Branching process dies out.
- $\eta(T,l) \leq 3\,(1 - 2\,\varepsilon)^l$
- Reconstruction requires $k \geq 2^{-7}\,(1 - 2\,\varepsilon)^{-l} = n^{\alpha}$
- Question: What if If $\theta(e) > \frac{1}{2} + \varepsilon$ for all e?

"random cluster" model summary

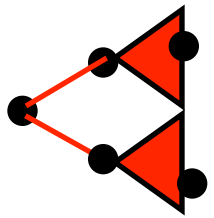
- Thm (Galton Watson): if $\theta(e) > \frac{1}{2} + \varepsilon$ for all e, then
  - for all $v \in T$,
  - with probability at least $s(\varepsilon) = 2\varepsilon / (\frac{1}{2} + \varepsilon)^2$,
  - there exists a leaf u which is a descendant of v, with $\sigma(v) = \sigma(u)$.
- Proof sketch:
  - Let $X(n)$ be the number of descendants u of v with $\sigma(v) = \sigma(u)$ where u is at distance n from v. Let $q(n) = P[X(n) = 0]$, and note that $q(0) = 0$, since $\sigma(v) = \sigma(v)$.
  - $q(n+1) \leq (1- p\,(1-q(n-1)))^2$, where $p = \frac{1}{2} + \varepsilon$.
  - Solving the recursion we see that there exists a descendent u of v with $\sigma(v) = \sigma(u)$ where u is at any distance from v is at least
  - $s(\varepsilon) = 2\varepsilon / (\frac{1}{2} + \varepsilon)^2$.

"random cluster" model summary

- Summary: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then:
- Branching process dies out.
- $\eta(T,l) \leq 3\,(1 - 2\,\varepsilon)^l$
- Phylogenetic reconstruction requires $k \geq 2^{-7}\,(1 - 2\,\varepsilon)^l$.
- Question: What if If $\theta(e) > \frac{1}{2} + \varepsilon$ for all e?
- Branching process does not die out.
- Is $\eta(T,l)$ decaying with l?

"random cluster" model summary

- Is $\eta(T,l)$ decaying with l?
- Claim: If $\theta(e) > \frac{1}{2} + \varepsilon$ for all e, then $\eta(T,l) \geq (\frac{1}{2} + \varepsilon)^2 s^2(\varepsilon)$
- Moreover with prob. at least $(\frac{1}{2} + \varepsilon)^2 s^2(\varepsilon)$ it is possible to recover the root color from the leaves.
- Proof (M-Steel-04)
  - If there are two leaf descendants u and w of v with the same color as v such that the only path from u to w is through v, then v must have the same color as u and w.



"Random cluster" model summary

- Summary: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then:
- Branching process dies out.
- $\eta(T,l) \leq 3\,(1 - 2\,\varepsilon)^l$
- Phylogenetic reconstruction requires $k \geq 2^{-7}\,(1 - 2\,\varepsilon)^l$.
- Question: What if If $\theta(e) > \frac{1}{2} + \varepsilon$ for all e?
- Branching process does not die out.
- Is $\eta(T,l)$ decaying with l? No $(\frac{1}{2} + \varepsilon)^2 s^2(\varepsilon)$
- Are polynomially many samples needed? Coming soon …

The CFN models on trees

- Most well knows is the Ising-CFN model on {-1,1}:

$$M^e = \begin{pmatrix} \frac{1+\theta(e)}{2} & \frac{1-\theta(e)}{2} \\ \frac{1-\theta(e)}{2} & \frac{1+\theta(e)}{2} \end{pmatrix}.$$

- Q: Assume that T is a balanced binary tree and that $\theta(e) = \theta$ for all e. How small is $\eta(T,l)$?
- A: This was studied intensively in Statistical physics under various names for balanced trees with fixed θ.
- It is known that if $2\theta^2 > 1$ then $\eta(T,l) > c(\theta) > 0$.
- It is known that if $2\theta^2 < 1$ then $\eta(T,l) \leq 0.5^{c(\theta)l}$
- $\Rightarrow$ Phylogeny recon. requires $k \geq n^c$

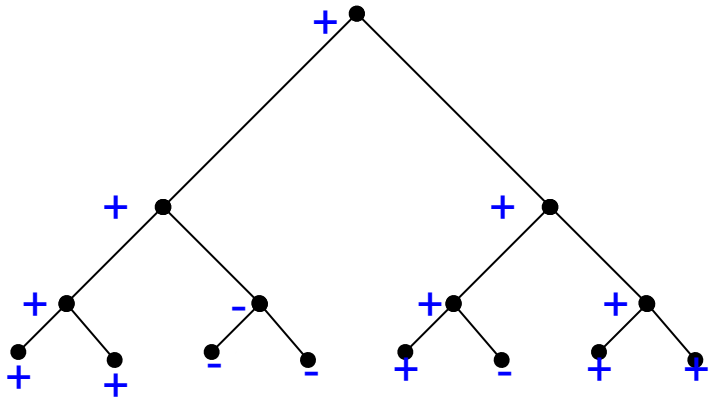
The CFN models on trees

- Most well knows is the Ising-CFN model on {-1,1}:
- Q: Assume that T is a balanced binary tree and that $\theta(e) = \theta$ for all e. How small is $\eta(T,l)$?
- A: This was studied intensively in Statistical physics under various names.
- It is know that if $2\theta^2 > 1$ then $\eta(T,l) > c(\theta)$.
- It is known that if $2\theta^2 < 1$ then $\eta(T,l)$ decays exponentially in l
- $\Rightarrow$ Phylogeny recon. requires $k \geq n^{\alpha}$

- Q: Suppose $2\theta^2 > 1$. Is it possible to reconstruct phylogenies with sequence length smaller than any polynomial?

The Ising model on the binary tree

- The (Free)-Ising-Gibbs measure on the binary tree:
- Set σ_r, the root spin, to be +/- with probability $\frac{1}{2}$.
- For all pairs of (parent, child) = (v, w), set $\sigma_w = \sigma_v$, with probability θ, otherwise σ_w = +/- with probability $\frac{1}{2}$.
- Different Perspective: Topology is known and looking at a single sample.



The Ising model on the binary tree

- Studied in **statistical physics** [Spitzer 75, Higuchi 77, Bleher-Ruiz-Zagrebnov 95, Evans-Kenyon-Peres-Schulman 2000, M 98]
- Interesting phenomena: double phase transition (different from Ising model in Z^d).
- When $2\theta^2 > 1$, unique Gibbs measure.
- When $2\theta^2 > 1$, free measure is extremal.
- In other words,

The Ising model on the binary tree

From BRZ or EKPS:

mutual information:
$H(\sigma_\partial) + H(\sigma_r) - H(\sigma_r,\sigma_\partial)$

Temp	θ	$\sigma_r \mid \sigma_\partial \equiv 1$	Uniq	$I(\sigma_r,\sigma_\partial)$	Free measure
high	$< 1/2$	unbiased	V	$\to 0$	extremal
med.	$(1/2,1/\sqrt{2})$	biased	X	$\to 0$	extremal
low	$> 1/\sqrt{2}$	biased	X	Inf > 0	Non-ext

Remark: $2\theta^2 = 1$ phase transition also transition for mixing time of *Glauber dynamics* for Ising model on tree (Berger, Kenyon, M, Peres)

What about other Markov models?

- In general not known. Some suggested readings:
- M-2001: First results showing "non-spectral behavior".
- M-Peres 2001, Janson-M 2004: "Census reconstruction" and "robust" reconstruction *are* determined by the second eigenvalue.
- Results for asymmetric CNF Model (Borgs-Chayes-M-Roch).
- Results for symmetric models on [q] (Sly 2008)
- Hard-Core models (Bhatnagar Sly Tetali)
- Recent connections to diluted spin-glasses (Mezard and Montanari 06)

Insertions and Deletions on Trees

- An important case that is not even approximately known is the case of insertions and deletions.
- Even the answer to the following is not known.
- Consider a mutation model where each letter is deleted with prob. p independently.
- Let $x \neq y$ two sequences of length $\leq n$ and let D_x (D_y) be the prob on sequence generated from x (y).
- How small can $|D_x - D_y|$ be?
- E.g: Can it be as small as $O(n^{-3})$?

ACGACCGTTGACCGACCCGACATTGTAAACTGT Original Sequence

ACGACCGTTGACCGACCCGACATTGTAAACTGT Deletions

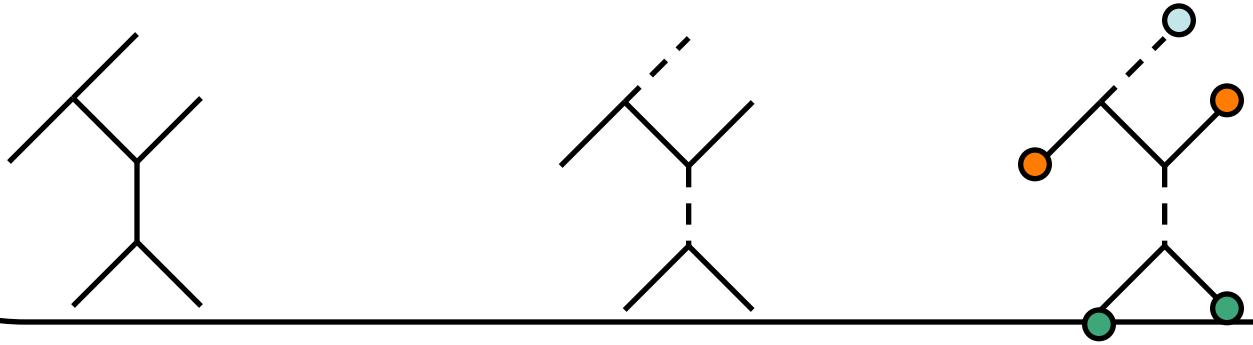
ACGCCGTTGACCGCCCGACTTGTAACTGT Mutated Sequence

Can we do better? Information Decay

- Conclusion of last lecture: Impossible to reconstruct if $k \leq 0.5 \log n$ and possible if $k \geq n^{\alpha}$. What is the truth?
- Last lecture we showed how if correlation decay holds then a polynomial lower bound holds.
- In this lecture we will ask if long range correlation (which is the opposite of correlation decay) hold phylogenies can be reconstructed from smaller # of samples.

The "random cluster" model

- Infinite set A of colors.
  - "real life" – large $|A|$; e.g. gene order.
- Defined on an un-rooted tree $T=(V,E)$.
- Edge e has (non-mutation) probability $\theta(e)$.
- Character: Perform percolation – edge e open with probability $\theta(e)$.
- All the vertices v in the same open-cluster have the same color σ_v. Different clusters get different colors. This is the "random cluster" model (both for (P,V,E) and $(P^{\otimes k}, V, E)$



"random cluster" model reconstruction

- Claim: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then the probability that
- there exists a leaf u which is a descended of v, with $\sigma(v) = \sigma(u)$ is at most $3\,(1 - 2\,\varepsilon)^{d(v,L)}$
- where d(v,L) is the tree distance between v and the leaf closest to v.
- Cor: Suppose $\theta(e) = \frac{1}{2} - \epsilon$ for all e and that T is a balanced binary tree of l levels rooted at r then:
- $\eta(T,l) \leq 3\,(1 - 2\,\varepsilon)^{l}$
- Pf Sketch: Let Q be the RC conditioned on having no path from the root to any of the leaves.
- Cor: The Phylogenetic reconstruction problem requires $k \geq 2^{-7}\,(1 - 2\,\varepsilon)^{-l}$ = 2^{-7} $n^{\log_2(1/1-2\,\varepsilon)}$ samples in order to reconstruct the tree with probability at least 0.9.

"random cluster" model summary

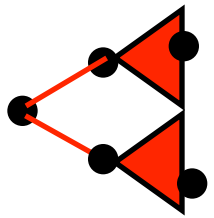
- Thm (Galton Watson): if $\theta(e) > \frac{1}{2} + \varepsilon$ for all e, then
  - for all $v \in T$,
  - with probability at least $s(\varepsilon) = 2\varepsilon / (\frac{1}{2} + \varepsilon)^2$,
  - there exists a leaf u which is a descendant of v, with $\sigma(v) = \sigma(u)$.
- Proof sketch:
  - Let $X(n)$ be the number of descendants u of v with $\sigma(v) = \sigma(u)$ where u is at distance n from v. Let $q(n) = P[X(n) = 0]$, and note that $q(0) = 0$, since $\sigma(v) = \sigma(v)$.
  - $q(n+1) \leq (1- p\,(1-q(n-1)))^2$, where $p = \frac{1}{2} + \varepsilon$.
  - Solving the recursion we see that there exists a descendent u of v with $\sigma(v) = \sigma(u)$ where u is at any distance from v is at least
  - $s(\varepsilon) = 2\varepsilon / (\frac{1}{2} + \varepsilon)^2$.

"random cluster" model summary

- Summary: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then:
- Branching process dies out.
- $\eta(T,l) \leq 3\,(1 - 2\,\varepsilon)^l$
- Phylogenetic reconstruction requires $k \geq 2^{-7}\,(1 - 2\,\varepsilon)^l$.
- Question: What if If $\theta(e) > \frac{1}{2} + \varepsilon$ for all e?
- Branching process does not die out.
- Is $\eta(T,l)$ decaying with l?

"random cluster" model summary

- Is $\eta(T,l)$ decaying with l?
- Claim: If $\theta(e) > \frac{1}{2} + \varepsilon$ for all e, then $\eta(T,l) \geq (\frac{1}{2} + \varepsilon)^2 s^2(\varepsilon)$
- Moreover with prob. at least $(\frac{1}{2} + \varepsilon)^2 s^2(\varepsilon)$ it is possible to recover the root color from the leaves.
- Proof (M-Steel-04)
  - If there are two leaf descendants u and w of v with the same color as v such that the only path from u to w is through v, then v must have the same color as u and w.



"Random cluster" model summary

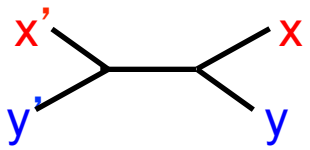
- Summary: If $\theta(e) < \frac{1}{2} - \varepsilon$ for all e, then:
- Branching process dies out.
- $\eta(T,l) \leq 3\,(1 - 2\,\varepsilon)^l$
- Phylogenetic reconstruction requires $k \geq 2^{-7}\,(1 - 2\,\varepsilon)^l$.
- Question: What if If $\theta(e) > \frac{1}{2} + \varepsilon$ for all e?
- Branching process does not die out.
- Is $\eta(T,l)$ decaying with l? No $(\frac{1}{2} + \varepsilon)^2 s^2(\varepsilon)$
- Are polynomially many samples needed?

Phylogeny from log characters for R.C.

- **Th1**[M-Steel,2004]: Suppose that T is a Phylogenetic tree on n leaves and for all e, $\frac{1}{2} + \varepsilon < \theta(e) < 1 - \varepsilon$.
  - Then $k = (2 \log n - \log \delta)/16\varepsilon^5 = O(\log n - \log \delta)$ characters suffice to reconstruct the topology with probability $\geq 1-\delta$.

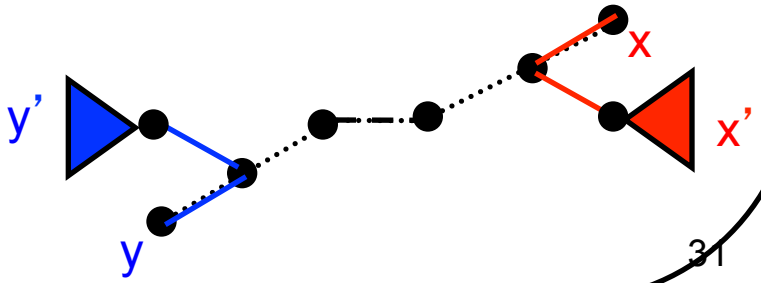
Testing cherries

- Claim: If x,y is a cherry **then** there exist **no** sample σ and leaves $x', y' \in \partial T - \{x,y\}$ s.t.
- $\sigma(x) = \sigma(x') \neq \sigma(y) = \sigma(y')$.



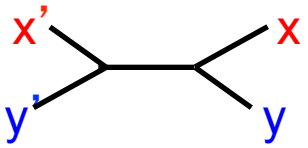
- Claim: If x,y is a not a cherry **then** for each sample σ,
  - $P[\exists x', y' \in L - \{x,y\}, \sigma(x) = \sigma(x') \neq \sigma(y) = \sigma(y')] \geq$
  - $P[\text{open edge}]^4 \times P[\text{closed edge}]$
    $\times P[v \in T, x \in L \text{ below } v, \sigma(x) = \sigma(v)]^2 \geq \epsilon\ s^2/16$,

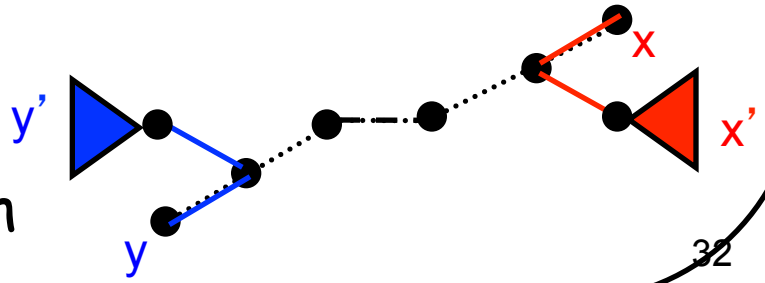
  where $s(\varepsilon) = 2\varepsilon / (\frac{1}{2}+\varepsilon)^2$.



Testing cherries

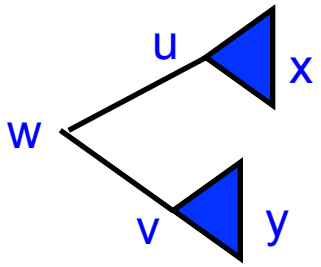
- We can repeat for k samples, looking for these witnesses.
- A pair of leaves that passed all tests will be declared a cherry.
- The only way the test fails is if a non-cherry pair of leaves has no witness in any of the k characters.
- So our probability of failure for each pair is bounded by $(1-r)^k$,
- giving us a total probability of failure bounded by $n^2 (1-r)^k$.
- With k = O(log n) samples can find all cherries with high prob.





From cherries to trees

- We wish to continue by replacing each cherry (u,v) by the parent w of v and u.
- Problem: We may not know what the color of w is.
- But: for each character σ, with probability at least $(\frac{1}{2} + \varepsilon)^2 s^2(e)$ we can reconstruct $\sigma(w)$.
- Now we can repeat.
- Result follows.



The CFN models on trees

- The Ising-CFN model on {-1,1}:

$$M^e = \begin{pmatrix} \frac{1+\theta(e)}{2} & \frac{1-\theta(e)}{2} \\ \frac{1-\theta(e)}{2} & \frac{1+\theta(e)}{2} \end{pmatrix}.$$

- Q: Assume that T is a balanced binary tree and that $\theta(e) = \theta$ for all e. How small is $\eta(T,l)$?
- A: This was studied intensively in Statistical physics under various names.
- It is know that if $2\theta^2 > 1$ then $\eta(T,l) > c(\theta)$.
- It is known that if $2\theta^2 < 1$ then $\eta(T,l)$ decays exponentially in l
- $\Rightarrow$ Phylogeny recon. requires $k \geq n^{\alpha}$

The CFN models on trees

- The Ising-CFN model on {-1,1}:
- Q: Assume that T is a balanced binary tree and that $\theta(e) = \theta$ for all e. How small is $\eta(T,l)$?
- A: This was studied intensively in Statistical physics under various names.
- It is know that if $2\theta^2 > 1$ then $\eta(T,l) > c(\theta)$.
- It is known that if $2\theta^2 < 1$ then $\eta(T,l)$ decays exponentially in l
- $\Rightarrow$ Phylogeny recon. requires $k \geq n^{\alpha}$

- Q: Suppose $2\theta^2 > 1$. Is it possible to reconstruct phylogenies with sequence length smaller than any polynomial?

The CFN models on trees

- Q: Suppose $2\theta^2 > 1$. Is it possible to reconstruct phylogenies with sequence length smaller than any polynomial?
- The answer to this question is quite involved.
- First we discuss an idealized case:
- Then talk about some more realistic models.
- Basic idea of reconstruction procedures is from M-2004 (which came before the simpler M-Steel-04)
- Iterate two steps:
- 1. Learn local tree structure
- 2. Reconstruct ancestral states.

Conjectures and results

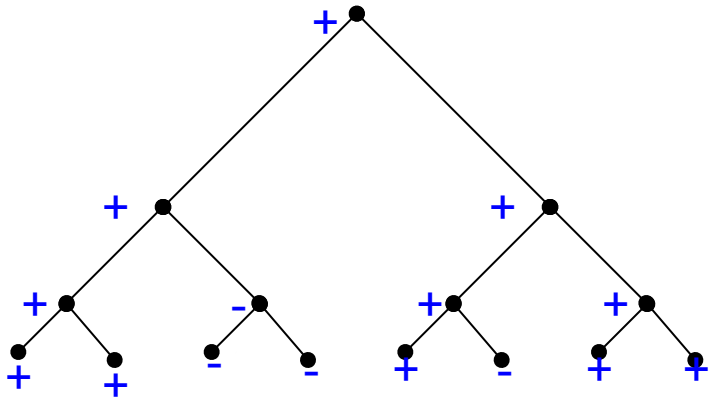
Statistical physics		Phylogeny
Binary tree in ordered phase	conj	k = O(log n)
Binary tree unordered	conj	k = poly(n)

Percolation	critical $\theta = 1/2$	Homoplasy free
Ising model	critical θ: $2\theta^2 = 1$	CFN

First conjecture is due to M. Steel 2001 for the CFN model

The Ising model on the binary tree

- Higuchi 77: Assume $2\theta^2 > 1$ and
- consider a binary tree of $l \geq 2$ levels and $\theta(e) = \theta$ for all e.
- Then $\mathrm{Cov}[\sigma(\mathrm{root})\ \mathrm{Maj}(\mathrm{leaves\ at\ level\ } l)] = \delta_l \geq \delta$.
- Pf (Exercise): Look at the sum conditioned on the root being + and calculate first and second moments.



An idealized Phylogenetic problem

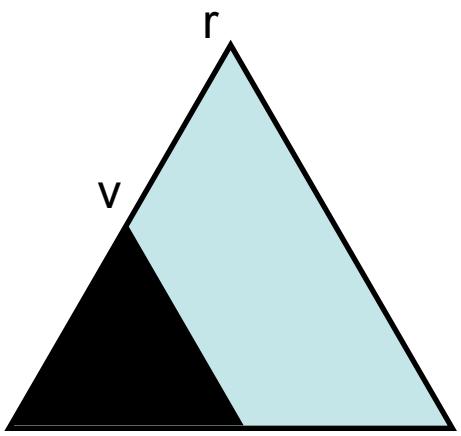
- Def: A tree is balanced if there exist a node r such that all leaves $x \in L$ are at the same graph distance R from r.
- Thm (M-2004): Suppose $2\,\theta^2 > 1$ then there exist an algorithm that requires $k = c(\theta,\delta) \log n$ samples that recovers a every possible Phylogenetic tree assuming:
- $\theta(e) = \theta$ for all e and
- the tree is balanced.
- (with error at most δ)

Algorithm sketch

- At iteration t of the algorithm we have disjoint balanced binary trees on $2t$ levels which cover all the leaves.
- Let u and v be two roots of such trees. Then:
- $E[Maj(L(T_u))\ Maj(L(T_v))] = E[\sigma(u)\sigma(v)] \times$
- $\times\ E[\sigma(u)\ Maj(L(T_u))]\ E[\sigma(v)\ Maj(L(T_v))]$
- $= E[\sigma(u)\sigma(v)]\ \ \delta(l)^2$
- We can therefore recover the sisters and cousins among u,v from $O(\log n)$ samples just by checking the correlation (using the fact that δ is bounded away from 0)

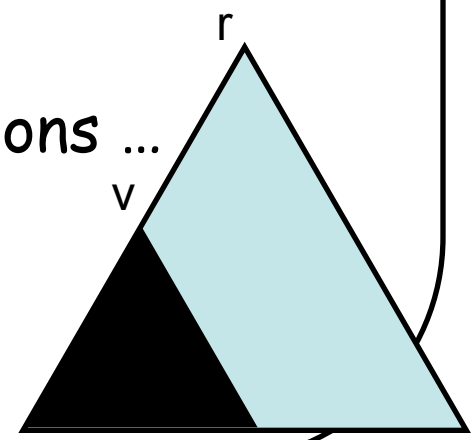
Algorithmic aspects of phase transition

- Can this be extended to the situation where there are different/unknown/approximate $\theta(e)$? The tree is not balanced?
- Looks good for phylogeny because can apply Maj even when do not know the topology.
- But, doesn't work when θ is non-constant.
  - All edges on blue subtree have θ^1
  - All edges on black subtree have θ^2
  - $\theta^1 < \theta^2$ is close to 1.
  - $Maj(\sigma_\partial)$ is very close to Maj of black tree.
  - Maj of black tree very close to σ_v.
  - σ_v and σ_r are weakly correlated.



Algorithmic aspects of phase transition

- Main idea of M-2004:
- If instead of Maj use Recursive-Maj then the problem "disappears":
- The correlation between the root and the leaves is high even if the $\theta(e)$ are non-homogenous.
- Allows to deal with either:
- $\theta(e)=\theta$ on all edges and general trees or
- Balanced tree and different θ values.
- Combining both in Daskalakis-M-Roch-10
- Unfortunately still some additional conditions …



More formal statement of lemma [M2004]

- Lemma: Assume that min $2\,\theta(e)^2 > 1 + \tau$.
- Then there exists an $l(\tau)$, and $\eta(\tau) > 0$ such that the CFN model on the binary tree of l levels with
  - $\theta(e) \geq \theta_{min}$, for all e not adjacent to ∂T.
  - $\theta(e) \geq \eta\,\theta_{min}$, for all e adjacent to ∂T.

  satisfies $E[\sigma(root)\,Maj(\sigma(L))] \geq \eta$.

- Roughly, given data of "quality $\geq \eta$", we can reconstruct the root with "quality $\geq \eta$".
- Does not require uniformity.
- Iterating the lemma gives that Rec-Maj is a good estimator.

More formal statement of lemma [M2004]

- Lemma: Assume that min $2\,\theta(e)^2 > 1 + \tau$.
- Then there exists an $l(\tau)$, and $\eta(\tau) > 0$ such that the CFN model on the binary tree of l levels with
  - $\theta(e) \geq \theta_{min}$, for all e not adjacent to ∂T.
  - $\theta(e) \geq \eta\,\theta_{min}$, for all e adjacent to ∂T.

  satisfies $E[\sigma(root)\,Maj(\sigma(L))] \geq \eta$.

- Pf sketch: The proof uses: isoperemetric inequalities, the random cluster representation etc. But some intuition can be gained from the case where η is small
- Linearize: $\eta_v = 0$ for all leaves but one.

More formal statement of lemma [M2004]

- Pf sketch: The proof uses: isoperemetric inequalities, the random cluster representation etc. But some intuition can be gained from the case where η is small
- Linearize: $\eta_v = 0$ for all leaves but w.
- Note: $\sigma(v)$ and $\sigma(r)$ independent for $v \neq w$
- $E[\sigma(w)\,\sigma(r)] \geq \theta^l\,\eta$
- $P[\mathrm{sgn}(\sum \sigma(v)) = \sigma(r)]$ "$\geq$" $(2/\pi)^{1/2}\, 2^{-l/2}\, \theta^l\, \eta$
- $\Rightarrow$ for small η_v's $\geq \eta$ we have:
- $P[\mathrm{sgn}(\sum \sigma(v)) = \sigma(r)]$ "$\geq$" $(2/\pi)^{1/2}\, 2^{l/2}\, \theta^l\, \eta$
- Obtain noise-reduction if: $2^{1/2}\,\theta > 1$.
- Easy to formalize but much more work is needed when some of the η_v's are large.

Algorithm sketch

- Similar to previous algorithm. Main difference in how to identify cherries.

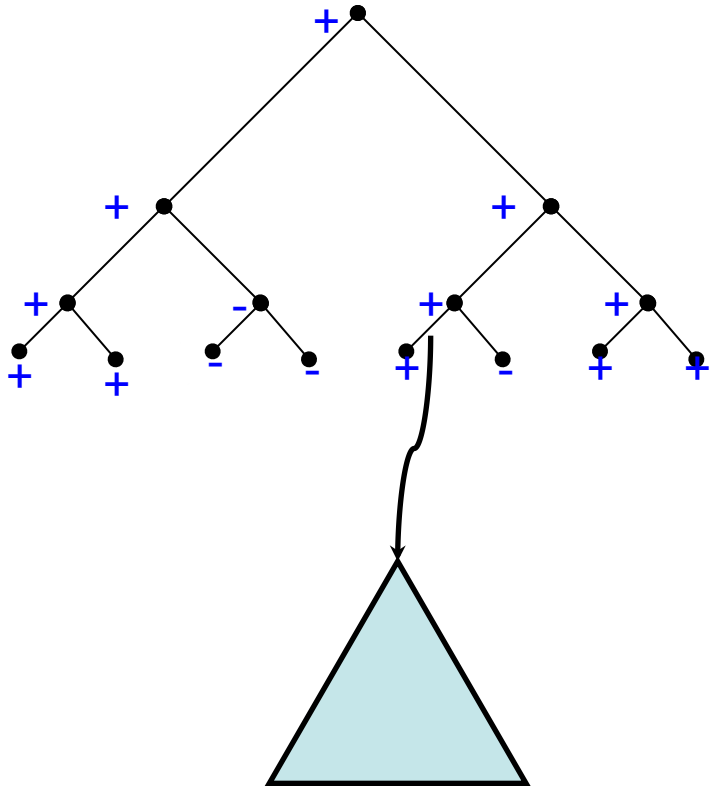
- If all θ's are the same use:
- $E[\text{R-Maj}(L(T_u))\ \text{R-Maj}(L(T_v))] = E[\sigma(u)\sigma(v)] \times$
- $\times\ E[\sigma(u)\ \text{R-Maj}(L(T_u))]\ E[\sigma(v)\ \text{R-Maj}(L(T_v))]$
- but last two terms depend on the shape of the tree.

- If tree is balanced but θ's are different also use four-point methods to identify cherries.

Some delicate combinatorial issues …

- In Daskalakis-M-Roch-10 need to deal with some delicate combinatorial issues.

- Note that if all $\theta(e)$ are the same we can recognize this in advance.
- If tree is balanced this will never happen.
- To control dependencies in DMR we require all edge length multiples of a small number.
- Open problem: remove this condition!



What about other Markov models?

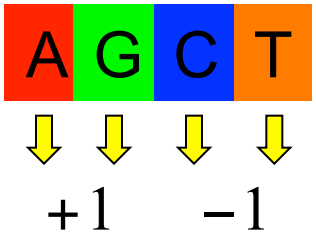
- The lower bounds holds for any Markov Model (M-Roch-Sly-10).
- For symmetric model (prob of mutation the same for any pair of states i,j):
- There are reconstruction algorithm with O(log n) samples if the mutation rate is below the Kesten-Stigum threshold (also known as the robust reconstruction, census reconstruction threshold).
- Even a little above the KS threshold (MRS-10)

A surprising result by Roch

- Is ancestral reconstruction really needed? Can we reconstruct from short sequences just from pairwise distances?
- assume $0 < \theta < \theta(e) << 1$, for all e where $2\theta^2 > 1$.
- Assume molecular clock (all leaves at the same metric distance from root)
- Thm (Roch-10): It is possible to reconstruct the tree from the empirical distances only given $O(\log n)$ samples!

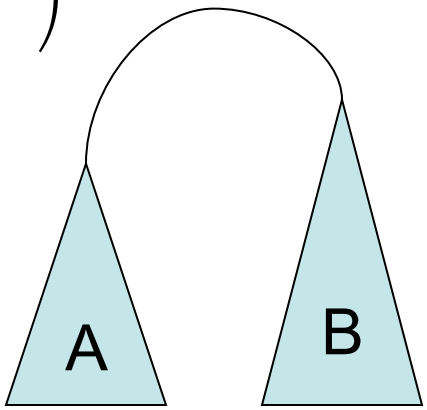
revisiting the averaging procedure I

- **step 1** - project the states to binary values



- the distance matrix becomes

$$D'(a,b) = -\ln\left(\frac{1}{k}\sum_{i=1}^{k} s_a^i s_b^i\right)$$



revisiting the averaging procedure II

- **step 2** – perform "exponential averaging" between clusters

$$D'(A,B) = -\ln\left(\sum_{a\in A}\sum_{b\in B} 2^{-|a|-|b|} e^{-D'(a,b)} \right)$$

$$= -\ln\left(\sum_{a\in A}\sum_{b\in B} 2^{-|a|-|b|} \frac{1}{k}\sum_{i=1}^{k} s_a^i s_b^i \right)$$

$$= -\ln\left(\frac{1}{k}\sum_{i=1}^{k} \left(\sum_{a\in A} 2^{-|a|} s_a^i \right)\left(\sum_{b\in B} 2^{-|b|} s_b^i \right) \right)$$

"majority"

$$D'(a,b) = -\ln\left(\frac{1}{k}\sum_{i=1}^{k} s_a^i s_b^i \right)$$