

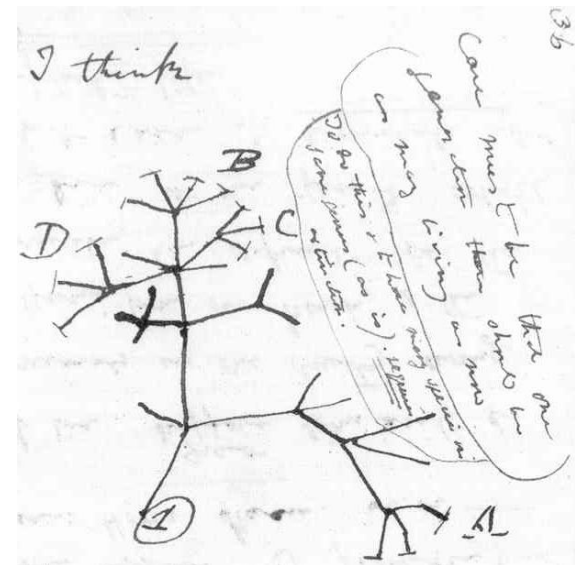
MARKOV MODELS ON TREES

reconstruction & applications

Sebastien Roch
Microsoft Research

collaborators:

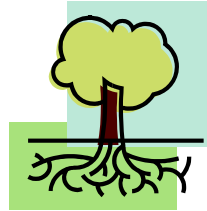
S. Bhamidi, C. Borgs, J. Chayes, C. Daskalakis,
E. Mossel, R. Rajagopal, M. Steel



outline of the talk

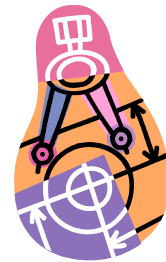
background & motivation

- the reconstruction problem
- main result
- motivation

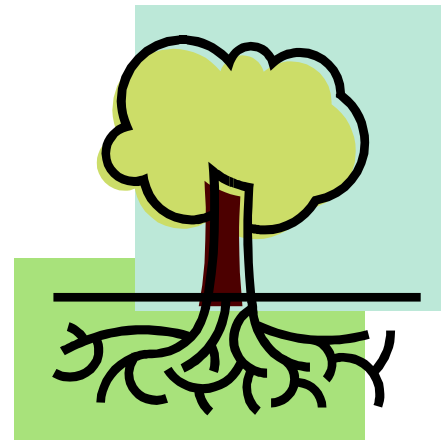


proof sketch

- magnetization
- recursions
- change of measure



PART I-a
background & main result



Markov chain on a tree

- broadcasting model

- b-ary tree: $T = (V, E)$
- node states:

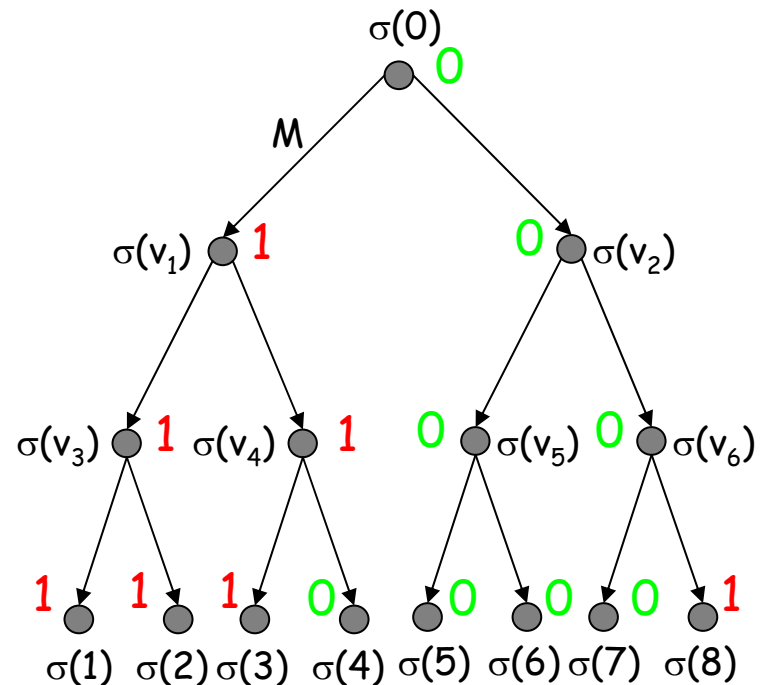
$$\{\sigma(v) \in \{0,1\} : v \in V\}$$

- Markov transition matrix:

$$M = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$$

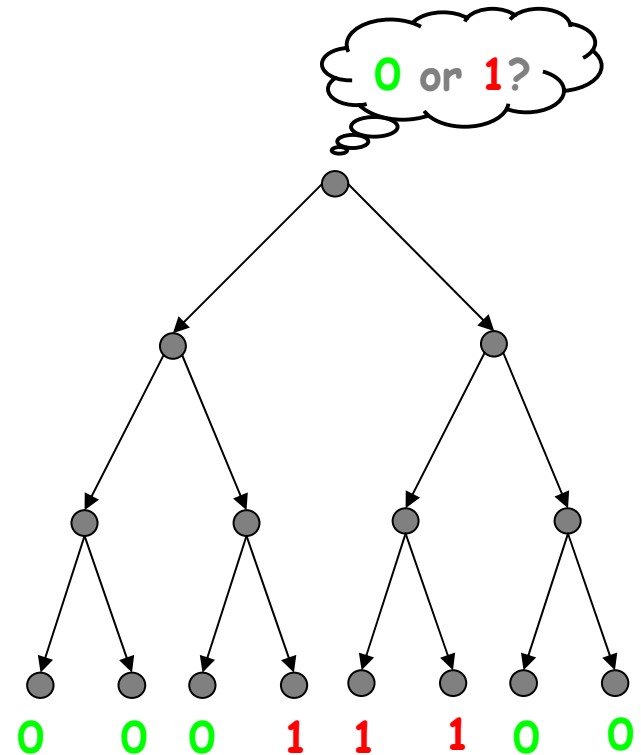
- stationary distribution:

$$\pi = (\pi_0, \pi_1)$$



the reconstruction problem

- ancestral reconstruction
 - **given**: states at leaves
 - **goal**: infer state at root
- phase transition
 - trade-off between **noise** and **duplication**



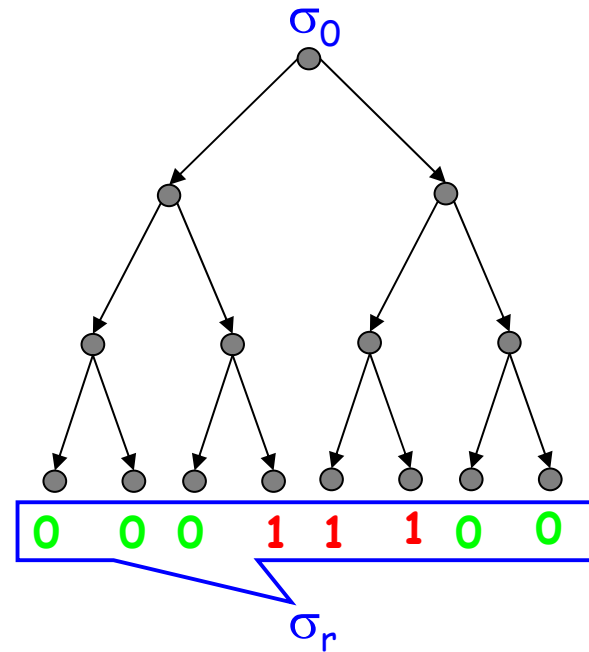
reconstruction solvability

- **setup**

- T : infinite rooted b -ary tree
- T_r : first r levels of T
- M : Markov transition matrix

- **definition** - the reconstruction problem on (T, M) is **solvable** if the following condition holds:

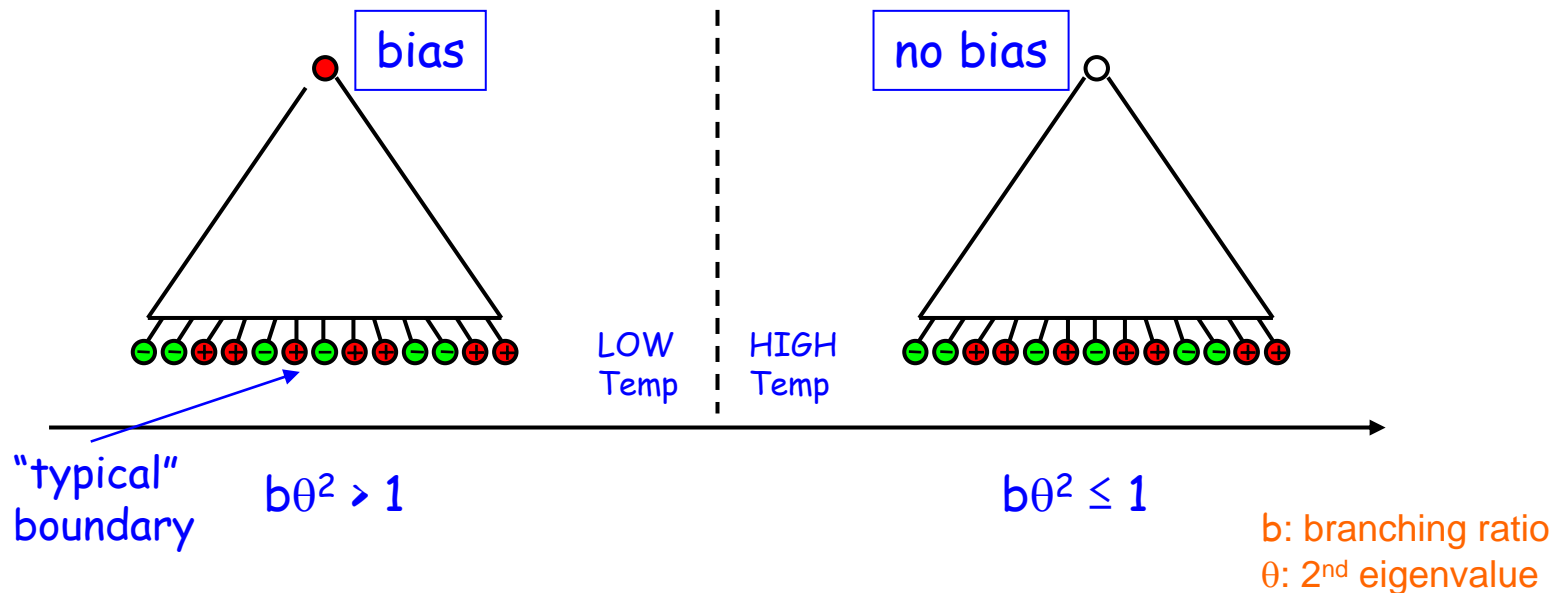
- $\lim_r |P_r^0 - P_r^1|_{TV} > 0$, where P_r^j denotes the distribution of σ_r conditional on $\sigma_0 = j$



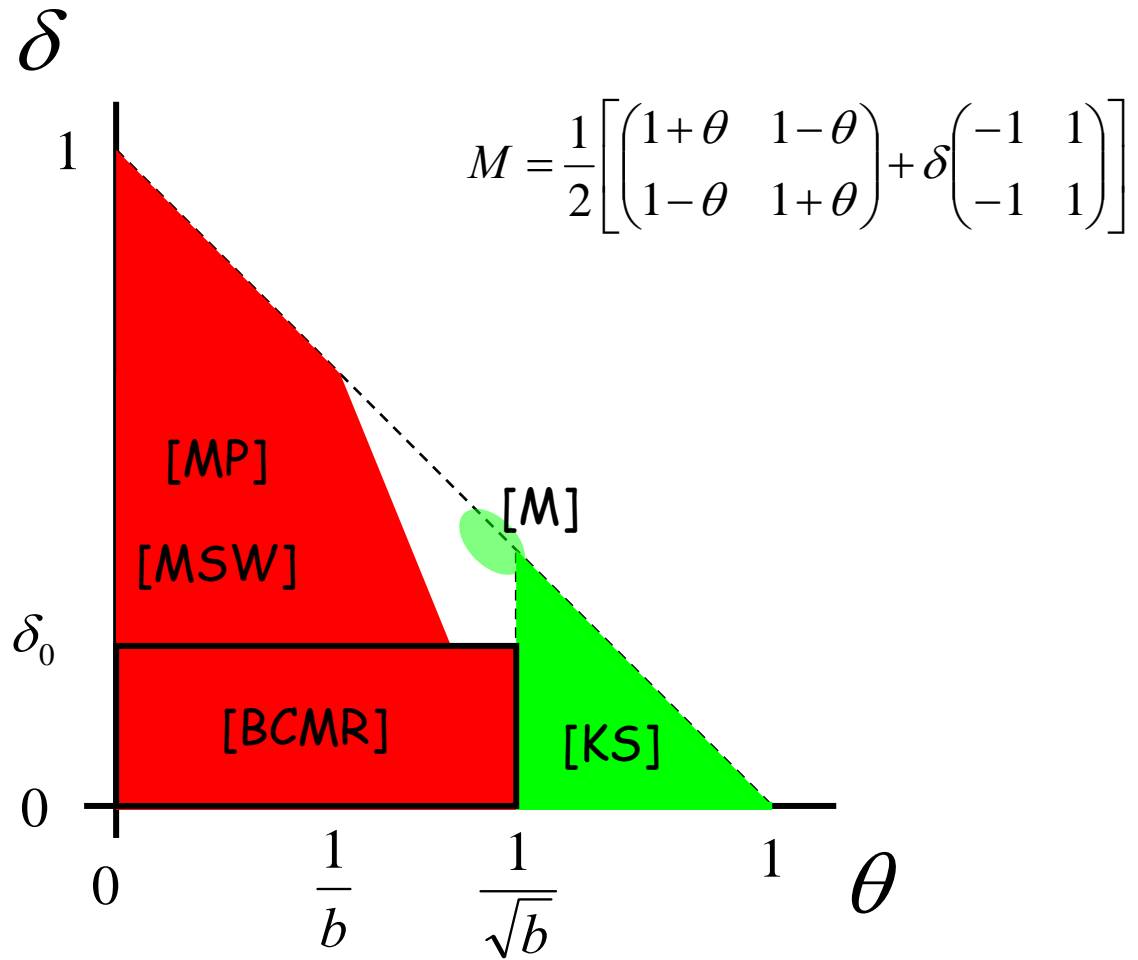
$$M = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \frac{1}{2} \left[\begin{pmatrix} 1+\theta & 1-\theta \\ 1-\theta & 1+\theta \end{pmatrix} + \delta \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \right], \text{ where } \theta = \lambda_2(M)$$

symmetric case ($\delta=0$)

- **theorem** - transition at $b\theta^2 = 1$
 - [Bleher-Ruiz-Zagrebnoy'95], [Ioffe'96], [Evans-Kenyon-Peres-Schulman'00], [Kenyon-Mossel-Peres'01], [Martin'03], [Martinelli-Sinclair-Weitz'04], [Borgs-Chayes-Mossel-R'06]
- solvability for $b\theta^2 > 1$ proved by [Higuchi'77], [Kesten-Stigum'66]
- "spinglass" case studied by [Chayes-Chayes-Sethna-Thouless'86]



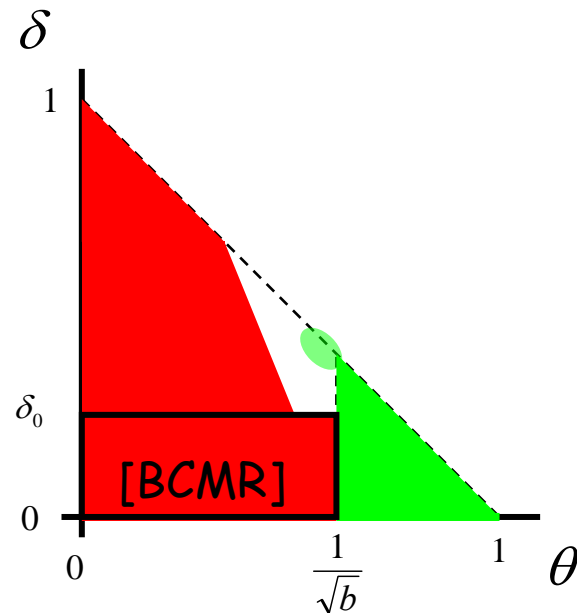
phase diagram



new result

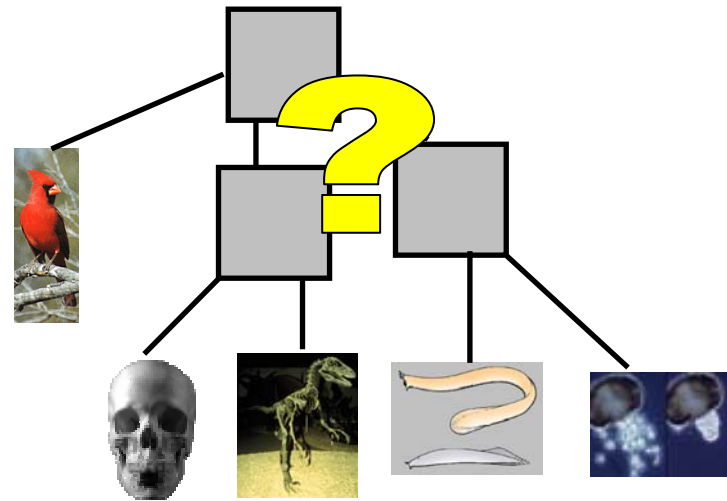
- **theorem** [Borgs-Chayes-Mossel-R'06] - exists $\delta_0 > 0$ s.t. if $b\theta^2 \leq 1$ and $|\delta| < \delta_0$ then the reconstruction problem is **not solvable**

- proof based on [Chayes-Chayes-Sethna-Thouless'86] who studied the "spinglass" case
- also new simple proof in symmetric case ($\delta = 0$)

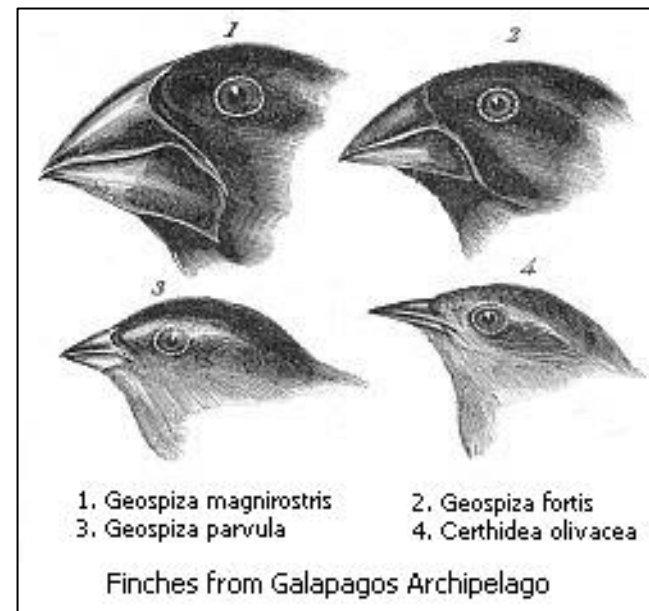
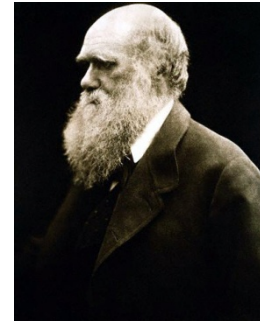


PART I-b

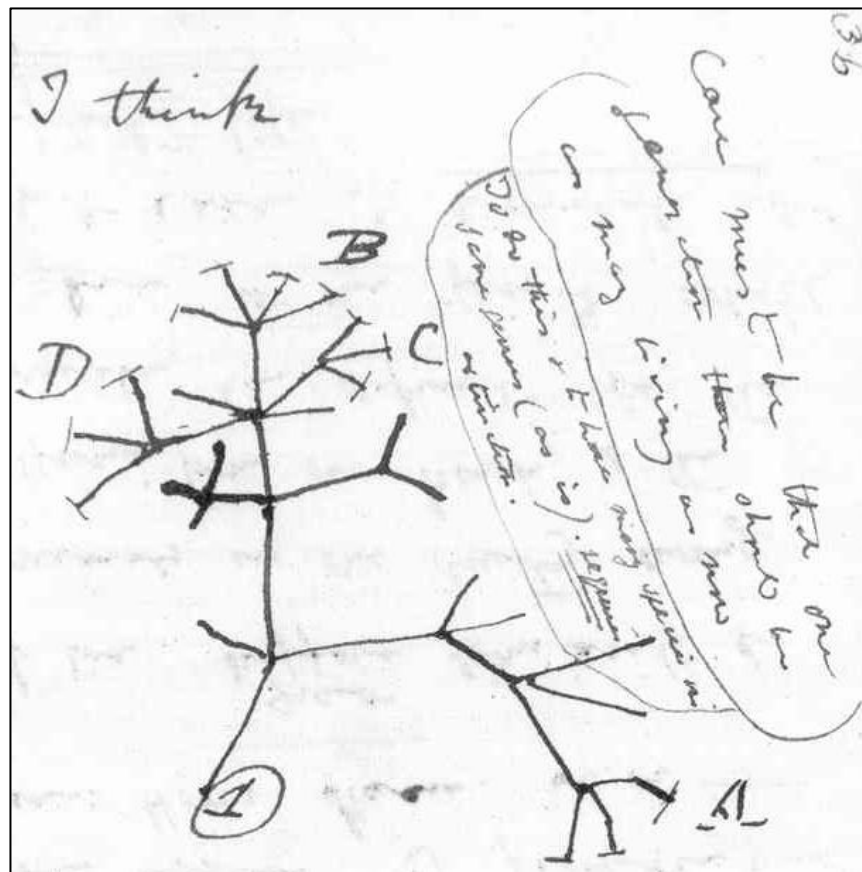
biological motivation



Darwin's finches



"i think"



From: Darwin, Transmutation Notebook B

DNA sequence evolution

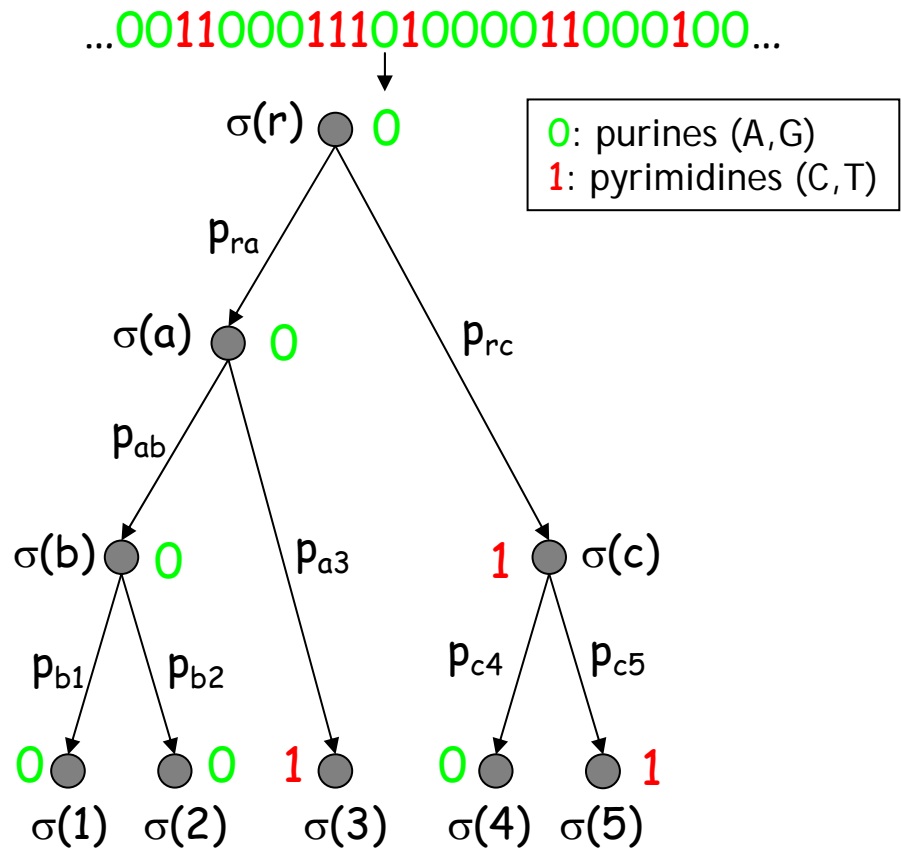
- CFN model

- only mutations
- tree: $T = (V, E)$
- node states:

$$\{\sigma(v) \in \{0,1\} : v \in V\}$$

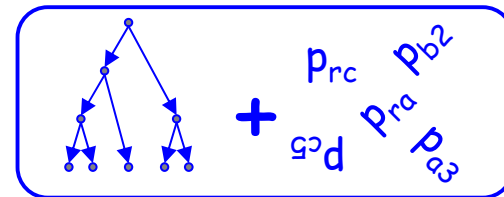
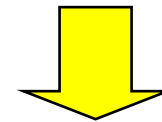
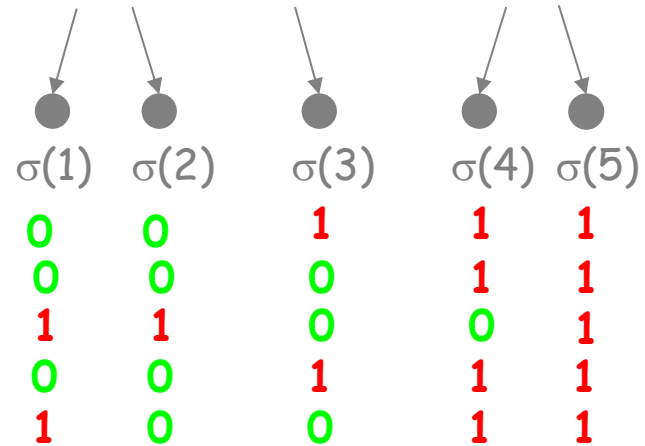
- mutation probabilities:

$$\{0 < p_e < 1/2 : e \in E\}$$



phylogenetic reconstruction

- phylogenetic reconstruction
 - **given**: sequences at the leaves
 - **goal**: fully reconstruct the model, i.e. find **tree** and **mutation** probabilities



more formally

- setup

- trees on n leaves: T_n
- model: $(T, \{p_e\}_{e \in E})$ in Θ_n
- k i.i.d. samples: $\sigma_L^1, \dots, \sigma_L^k$
- reconstruction map:

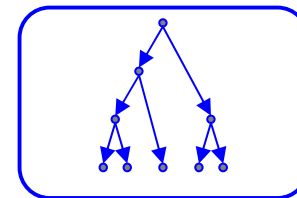
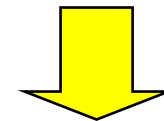
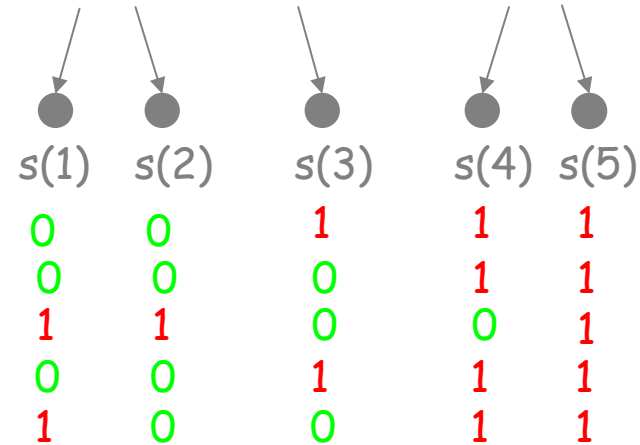
$$\Psi_n : \left\{ \sigma_L^i \right\}_{i=1}^k \mapsto T \in T_n$$

- **definition** - the map Ψ_n solves the **phylogenetic reconstruction problem** with k samples and confidence $1-\delta$ if for all models $(T, \{p_e\}_{e \in E})$ in Θ_n

$$P\left[\Psi_n\left(\left\{\sigma_L^i\right\}_{i=1}^k\right) = T\right] \geq 1 - \delta$$

- efficiency

- computational: running time
- information-theoretic: k



maximum likelihood

- **data:** n $\{0,1\}$ -sequences of length k

$$\{S(j) = (\sigma_L^1(j), \dots, \sigma_L^k(j)) \in \{0,1\}^k : 1 \leq j \leq n\}$$

- **likelihood**

$$\Lambda(T, \{p_e\}; S(1), \dots, S(n)) = \prod_{i=1}^k \sum_{\sigma^* \in \text{Ext}(\sigma_i)} \prod_{e=(u,v) \in E} p_e^{\langle \sigma^*(u) \neq \sigma^*(v) \rangle} (1 - p_e)^{\langle \sigma^*(u) = \sigma^*(v) \rangle}$$

- **MLE**

$$(T^*, \{p_e^*\}) = \arg \min_{(T, \{p_e\})} [-\ln \Lambda(T, \{p_e\}; S(1), \dots, S(n))]$$

- statistically consistent [Chang'96], but:
 - **theorem** [Chor-Tuller'05, R'06] - NP-hard (i.e. "computationally intractable"); actually hard to approximate

distance matrix methods

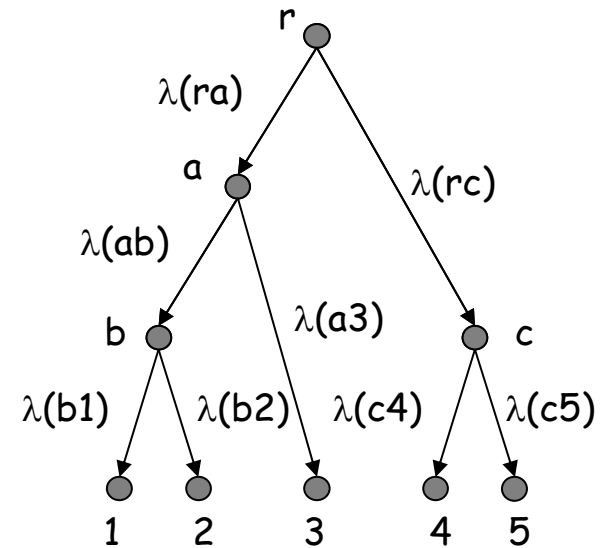
- in CFN (i.e. 0-1 symmetric) case:
 - associate to each edge e a **weight**

$$\lambda(e) = -\ln(1 - 2p_e)$$

- defines a **tree metric**

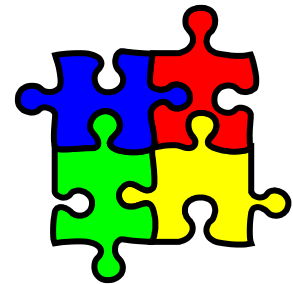
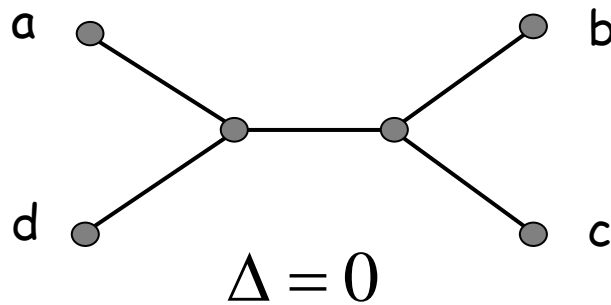
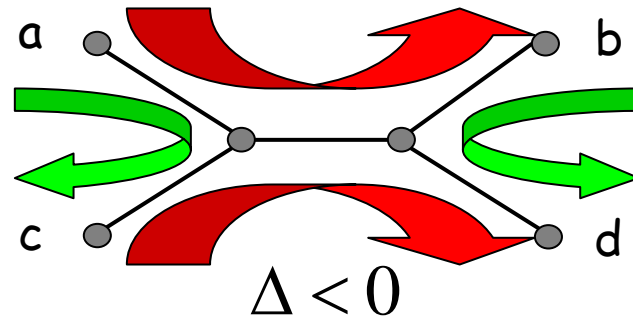
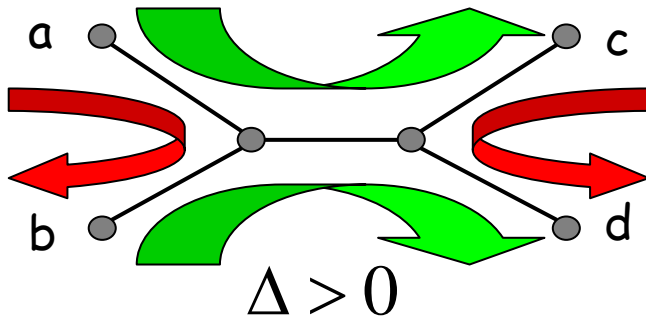
$$D(i, j) = \sum_{e \in P(T; i, j)} \lambda(e) = -\ln(1 - 2P[\sigma(i) \neq \sigma(j)])$$

- generalized by [Steel'94]
- reconstruction algorithm:
 - estimate $D(i, j)$ from sequences
 - deduce the topology of the tree
- **theorem** - reconstruction can be done efficiently (polynomial time)
 - e.g. Neighbor-Joining



four-point method

$$\Delta = D(a,c) + D(b,d) - D(a,b) - D(c,d)$$



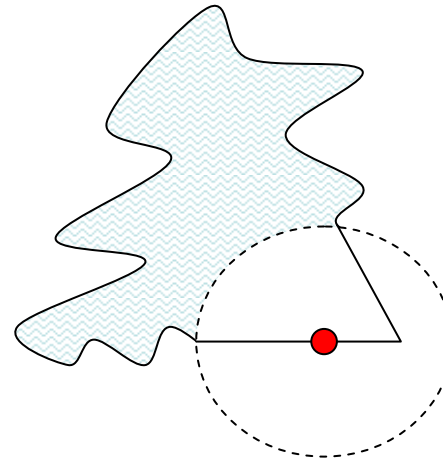
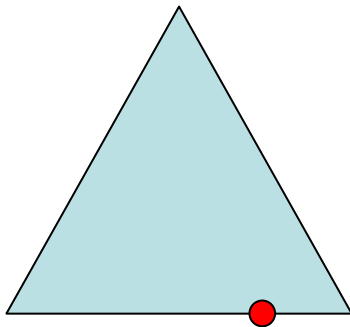
“local” metric

- **assumption (A)** - for simplicity assume $0 < f < \lambda(\mathbf{e}) < g$, for all \mathbf{e}
- **fact** - to estimate distances of order M with precision ε , one needs

$$k = \frac{C}{(1 - e^{-2\varepsilon})^2} e^{2M+2\varepsilon} \log n$$

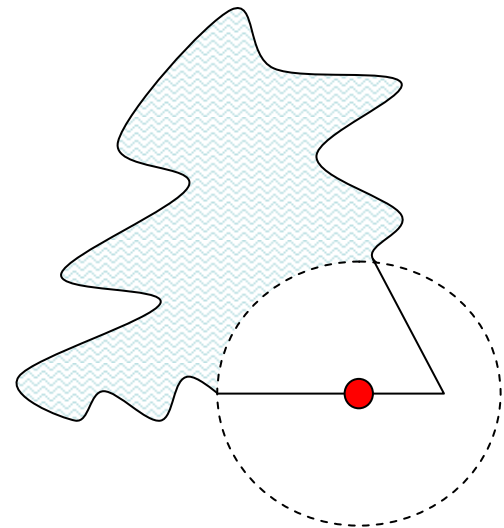
- **definition** [Mossel'07] - a symmetric matrix \mathbf{d} is a (ε, M) -**distortion** of the distance matrix \mathbf{D} if

$$|D(i, j) - d(i, j)| < \varepsilon \text{ if } d(i, j) < M + \varepsilon \text{ or } D(i, j) < M + \varepsilon$$



distance methods revisited

- **assumption (A)** - for simplicity assume $0 < f < \lambda(e) < g$, for all e
- **theorem** [Erdos-Steel-Szekely-Warnow'97,'98], [Mossel'07], [Daskalakis et al.'06], [Gronau-Moran-Snir'08], [Daskalakis-Mossel-R'07] - can achieve polynomial-length sequences and polynomial time

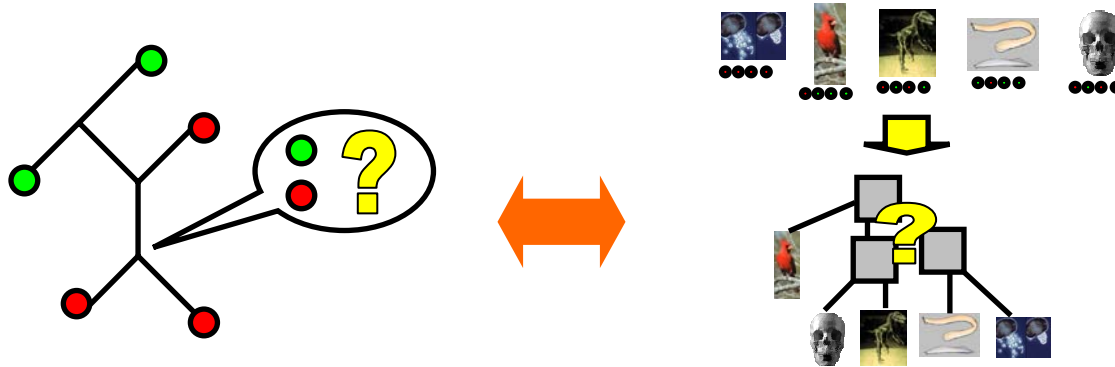


distance methods: recap

- **summary**: phylogenies can be inferred in polynomial time from polynomial length sequences
- **questions**: is this the best we can do?
- **counting argument**: need at least $\Omega(\log n)$ samples...

$$2^{O(n \log n)} \approx 2^{nk}$$

resolution of Steel's conjecture



ancestral
reconstruction

phylogenetic
reconstruction

[Daskalakis-
Mossel-R'06]

solvable



$k = O(\log n)$

[Mossel'04,
Borgs-
Chayes-
Mossel-R'06]

not solvable



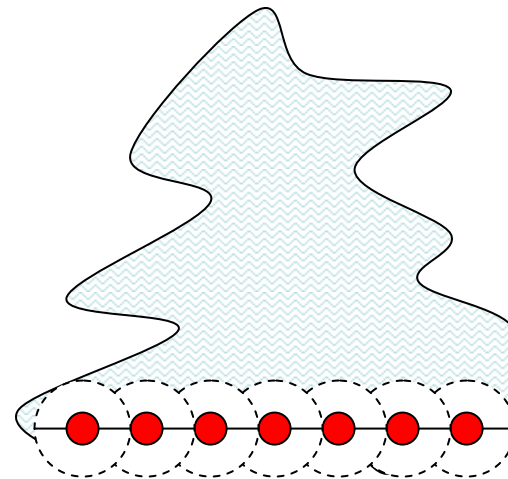
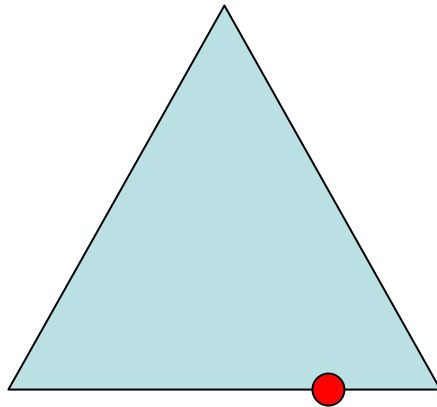
$k = \text{poly}(n)$

$k = \#$ of samples
 $n = \#$ of leaves

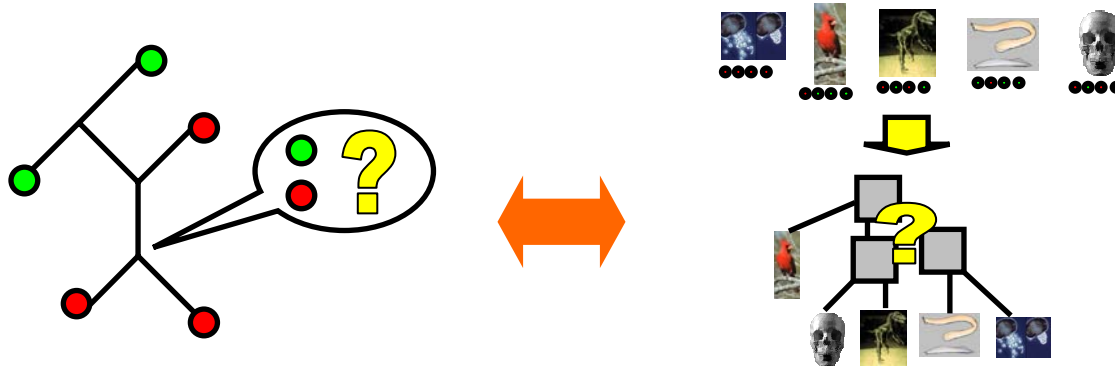
“very local” metric

- **recall** - to estimate distances of order M with precision ε , one needs

$$k = \frac{C}{(1 - e^{-2\varepsilon})^2} e^{2M+2\varepsilon} \log n$$



resolution of Steel's conjecture



ancestral
reconstruction

phylogenetic
reconstruction

[Daskalakis-
Mossel-R'06]

solvable



$k = O(\log n)$

[Mossel'04,
Borgs-
Chayes-
Mossel-R'06]

not solvable

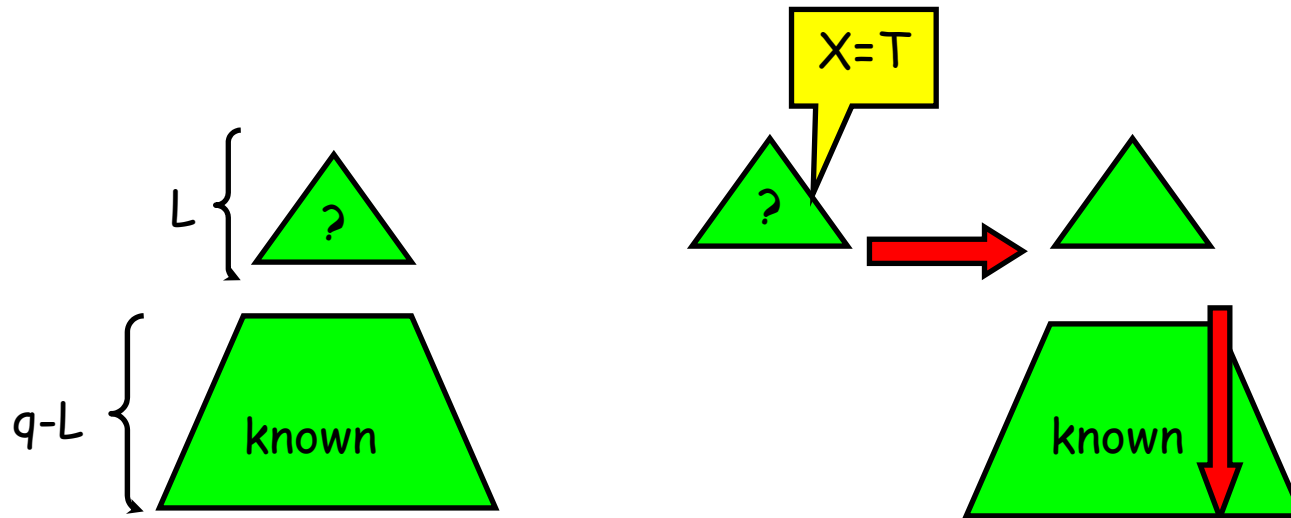


$k = \text{poly}(n)$

$k = \#$ of samples
 $n = \#$ of leaves

polynomial lower bound

- proof



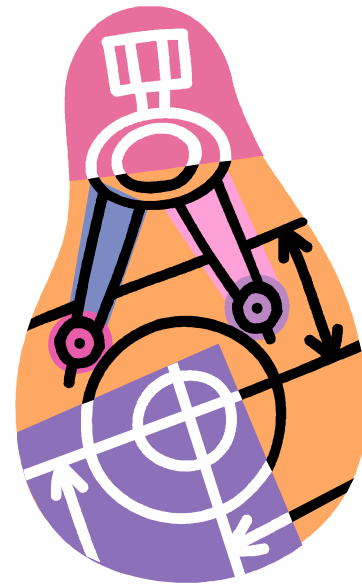
- **mutual information:** $I(X,Y) = H(X) - H(X | Y)$
- **data processing lemma:** if X and Z are cond. indep. given Y then $I(X,Y) \geq I(X,Z)$

PART II

proof sketch

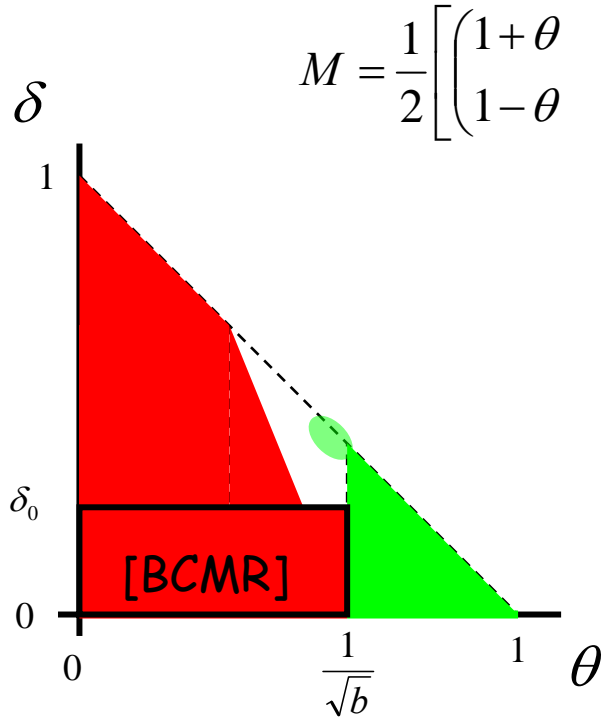
based on:

Borgs, Chayes, Mossel, R,
The Kesten-Stigum Bound is Tight
for Roughly Symmetric Channels,
Proceedings of IEEE FOCS'06

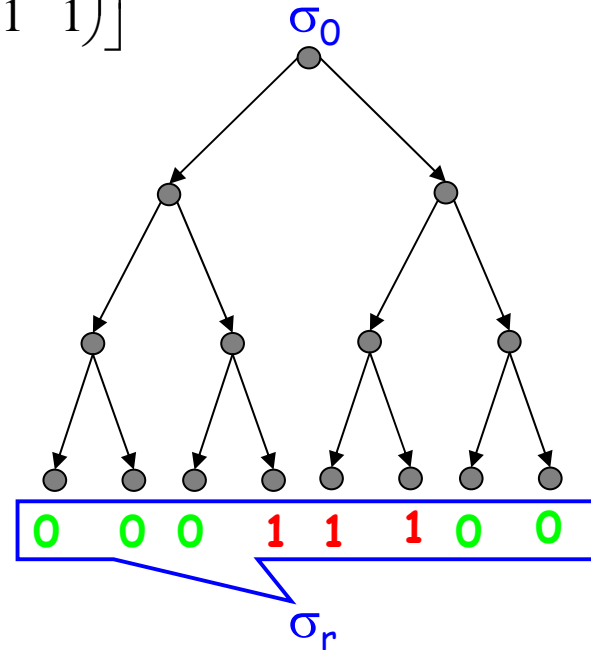


recall

- theorem** [Borgs-Chayes-Mossel-R'06] - exists $\delta_0 > 0$ s.t. if $b\theta^2 \leq 1$ and $|\delta| < \delta_0$ then the reconstruction problem is **not solvable**



$$M = \frac{1}{2} \left[\begin{pmatrix} 1+\theta & 1-\theta \\ 1-\theta & 1+\theta \end{pmatrix} + \delta \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix} \right]$$



magnetization of the root

- we use $\{+1, -1\}$.
- stationary distribution of channel \mathbf{M}

$$\pi = (\pi_+, \pi_-)$$

- **definition** - the magnetization of the root is

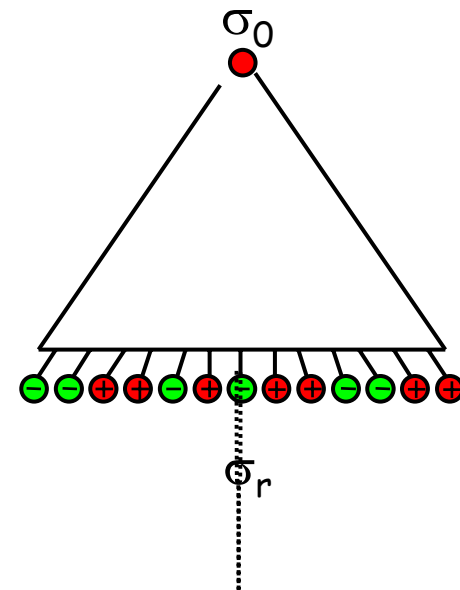
$$X_r(\sigma_r) = \pi_-^{-1} \{ \pi_- \text{P}[\sigma_0 = +1 | \sigma_r] - \pi_+ \text{P}[\sigma_0 = -1 | \sigma_r] \}$$

- **lemma** - it suffices to show

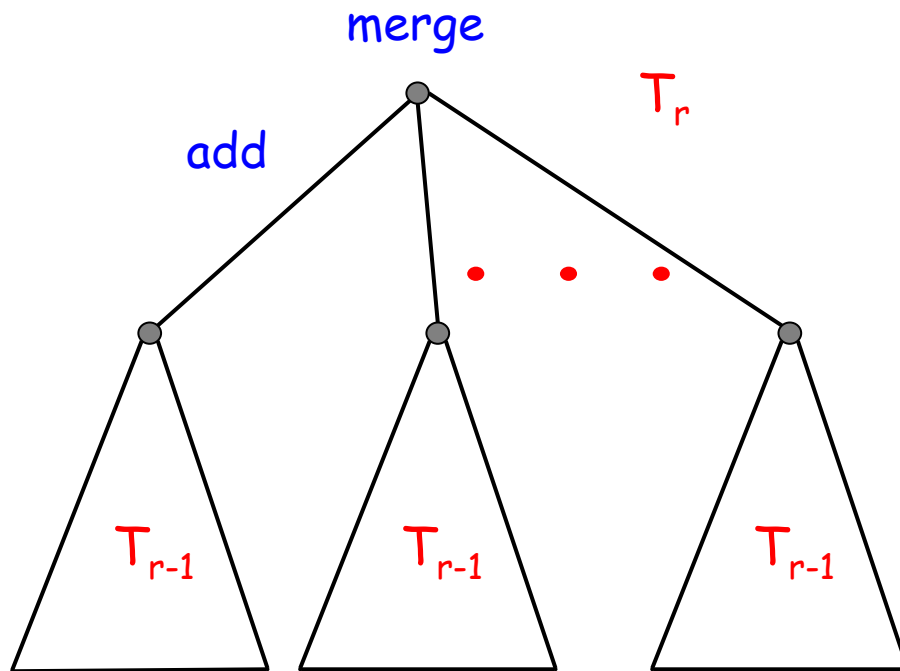
$$\bar{x}_r \equiv \mathbb{E}_{T_r}^{\pi} [X_r^2] \rightarrow 0$$

- basic proof idea: **moment recursion**

$$\bar{x}_r \leq b\theta^2 \bar{x}_{r-1}$$



from T_{r-1} to T_r



add-merge

- using **Bayes** and **Markov**

$$X = \frac{Y + \theta Z + (\pi_- \pi_+^{-1} - 1)\theta YZ}{1 + \pi_- \pi_+^{-1} \theta YZ}$$

- perform the **change of measure**

$$E_T^+[X] = E_T^\pi[X(1 + \pi_- \pi_+^{-1} X)] = \pi_- \pi_+^{-1} E_T^\pi[X^2]$$

- expand

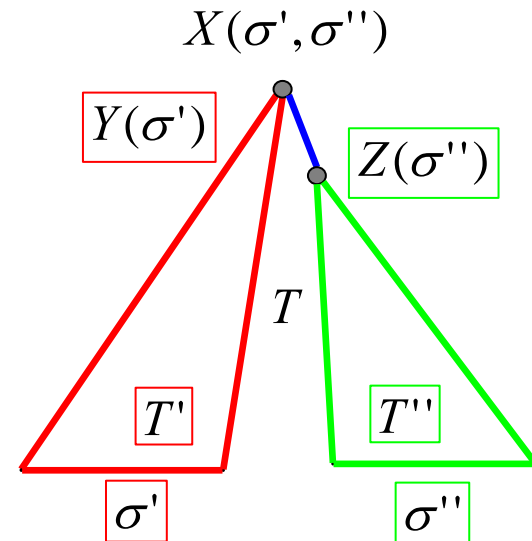
$$X \leq Y + \theta Z + (\pi_- \pi_+^{-1} - 1)\theta YZ + (\pi_- \pi_+^{-1} \theta YZ)^2 - \pi_- \pi_+^{-1} \theta YZ [Y + \theta Z + (\pi_- \pi_+^{-1} - 1)\theta YZ]$$

- we get the **basic inequality**

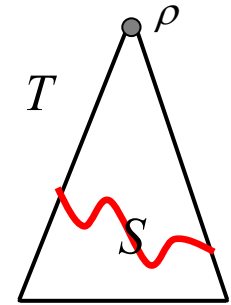
$$\bar{x} \leq \bar{y} + \theta^2 \bar{z}$$

- repeating **b-1** times

$$\bar{x}_r \leq b \theta^2 \bar{x}_{r-1}$$



more general result



- general trees - previous results

- **definition** - the **branching number** is defined as

$$\text{br}(T, \theta) = \inf \left\{ \lambda > 0 : \inf_{\text{cutsets } S} \sum_{x \in S} \left(\lambda^{-|x|} \prod_{e \in \text{path}(\rho, x)} \theta^2(e) \right) = 0 \right\}$$

- [Evans-Kenyon-Peres-Schulman'00] binary symmetric case on general tree, solvable iff $\text{br}(T, \theta) > 1$

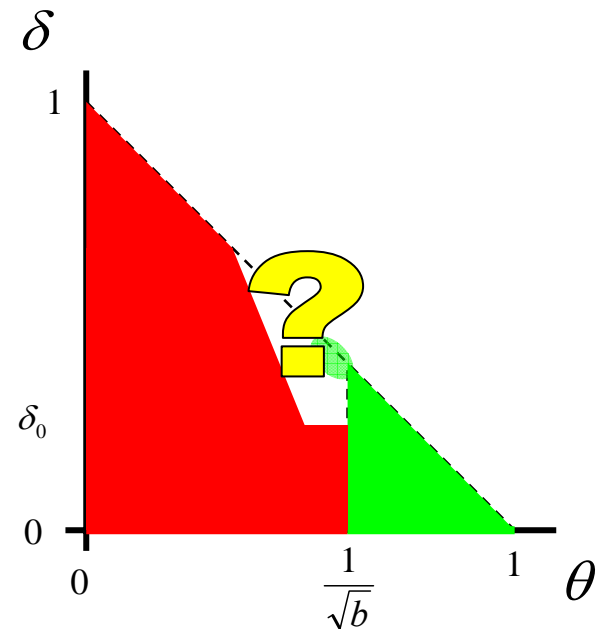
- **theorem** [Borgs-Chayes-Mossel-R'06] - let $0 \leq \theta_0 < 1$. exists $\delta_0 > 0$ such that

- for all stationary distributions $\pi = (\pi_+, \pi_-)$ with $\max\{|\delta(\pi, \theta)|, |\delta(\pi, -\theta)|\} < \delta_0$
- for all trees with $\sup_e |\theta(e)| \leq \theta_0$ and $\text{br}(T, \theta) \leq 1$

the reconstruction problem is **not solvable**

final remarks

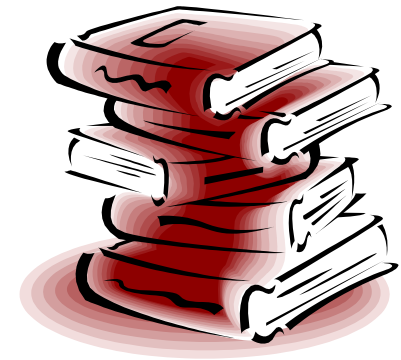
- complete phase diagram
- Potts model: conjectures by [Mezard-Montanari'06]
- positive result in asymmetric case
- random trees, mixture models
- population-genetic effects



thank
you

references

- R, "A Short Proof that Phylogenetic Tree Reconstruction by Maximum Likelihood is Hard", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2006
- Mossel, R, "Learning Nonsingular Phylogenies and Hidden Markov Models", *Annals of Applied Probability*, 2006
- Daskalakis, Mossel, R, "Optimal Phylogenetic Reconstruction", *Proceedings of ACM STOC'06*
- Borgs, Chayes, Mossel, R, The Kesten-Stigum Bound is Tight for Roughly Symmetric Channels, *Proceedings of IEEE FOCS'06*
- Daskalakis, Mossel, R, "Phylogenies Without Branch Bounds: Contracting the Deep, Pruning the Deep", *Preprint*, 2007



phase diagram

