Group Meeting - September 24, 2021

Truong Son Hy *

*Department of Computer Science
The University of Chicago

Ryerson Physical Lab

Paper

Tropical geometry of statistical models
Lior Pachter and Bernd Sturmfels
`https://www.pnas.org/content/101/46/16132`

Note:

- The paper is very dense and contains a number of typos and a number of flows.
- I don't understand everything in it and decided to present one flow (that I find the most meaningful).
- I keep my comments and fixes in blue color.

Algebraic variety & Tropicalization

- An algebraic variety is defined as the **set of solutions** of a system of polynomial equations over the real or complex numbers.
- Tropicalization means replacing the arithmetic operations $(+, \times)$ by the operations $(\min, +)$:

$$a + b \to \min\{a, b\}$$

$$a \times b \to a + b$$

For example, polynomial

$$a^3 + 2ab + b^2$$

would become

$$\min\{a + a + a, 2 + a + b, b + b\}.$$

Indeed, the tropicalization of an algebraic variety is a **piecewise-linear set**.

Proposal

A **unified** mathematical framework for probabilistic inference with statistical models:

1. Statistical models are algebraic varieties. Families of joint probability distributions $p_{\sigma_1,..,\sigma_n} = P(Y_1 = \sigma_1, .., Y_n = \sigma_n)$ can be characterized by polynomials.
   This seems to be trivial to me, counter-example?
2. Every algebraic variety can be tropicalized. The joint probabilities $p_{\sigma_1,..,\sigma_n}$ are replaced by their logarithms (i.e., multiplication turns into sum).
   Literally, I think what they mean is: the solution or algebraic variety is expressed in terms of $(\min, +)$. For example, the Viterbi algorithm (dynamic programming) is exactly this.
3. Tropicalized statistical models are fundamental for parametric inference.
   I think parametric inference is an abstraction over existing inference algorithms.

Directed graphical model (1)

A directed graphical model (or Bayesian network) is a finite directed acyclic graph G with two kinds of vertices:

- **Observed** variables $\boldsymbol{Y} = \{Y_1, .., Y_n\}$,
- **Hidden** variables $\boldsymbol{X} = \{X_1, .., X_m\}$,

where each edge is labeled by a transition matrix whose entries are linear forms in some parameters. The observed probabilities $p_{\sigma_1,..,\sigma_n}$ is expressed as polynomials of a degree $\leq E$ in the parameters, where E is the number of edges of G.

Directed graphical model (2)

Two types of inference questions from statistical learning theory for graphical models:

1. The calculation of **marginal probabilities**:

$$p_{\sigma_1,..,\sigma_n} = \sum_{h_1,..,h_m} P(X_1 = h_1, .., X_m = h_m, Y_1 = \sigma_1, .., Y_n = \sigma_n),$$

2. The calculation of **maximum a posteriori** (MAP) log probabilities:

$$\delta_{\sigma_1,..,\sigma_n} = \min_{h_1,..,h_m} -\log P(X_1 = h_1, .., X_m = h_m, Y_1 = \sigma_1, .., Y_n = \sigma_n),$$

where h_i range over all of the possible assignments for the hidden random variables X_i.

Directed graphical model (3)

About sum-product algorithm, message passing, belief propagation (exact algorithm on trees, acyclic graphs and approximate on other topologies) on graphical models, please see August 14th, 2020 presentation:

`http://people.cs.uchicago.edu/~hytruongson/Discussions-2020/Group_meeting___August_14__2020.pdf`

Algebraic representation of HMMs (1)

A graphical model is an algebraic variety that is presented as the image of a highly structured polynomial map $f : \mathbb{R}^d \to \mathbb{R}^m$.

- $\mathbb{R}^d$ is the space in which coordinates are the model parameters $s_1, .., s_d$.
- $\mathbb{R}^m$ is the space in which coordinates $p_\sigma = p_{\sigma_1,..,\sigma_n}$ are the joint probabilities for the observed random variables.
- Each coordinate $f_\sigma = f_\sigma(s_1, .., s_d)$ of the map f is a polynomial function in $s_1, .., s_d$.

Important note

- The efficient evaluation of these functions relies on the sum-product algorithm.
- But **parametric** inference shows/analyzes the relationship between model parameters vs. values of hidden states.

Algebraic representation of HMMs (2)

A discrete HMM has n observed states $Y_1, .., Y_n$ taking on ℓ possible values, and n hidden states $X_1, .., X_n$ taking on k possible values. The HMM can be characterized by the following conditional independence statements for $i = 1, .., n$:

$$P(X_i|X_1, .., X_{i-1}) = P(X_i|X_{i-1}),$$

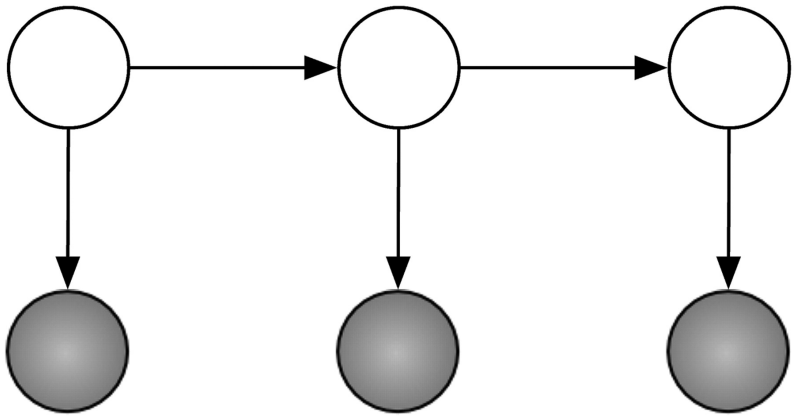
$$P(Y_i|X_1, .., X_n, Y_1, .., Y_{i-1}) = P(Y_i|X_i).$$

All transitions $X_i \to X_{i+1}$ are given by the same $k \times k$ matrix $S = (s_{ij})$.
All transitions $X_i \to Y_i$ are given by the same $k \times \ell$ matrix $T = (t_{ij})$.

Algebraic representation of HMMs (3)

Proposition 1. *The HMM is the image of a map* $f\colon \mathbf{R}^d \to \mathbf{R}^{l^n}$, *where* $d = k(k+1)$ *and each coordinate of* f *is a bihomogeneous polynomial of degree* $n-1$ *in* S *and degree* n *in* T.

I think there is a typo here: $d = k(k+\ell)$.



The parameter space is $\mathbb{R}^8$ with the coordinates:

$$s_{00}, s_{01}, s_{10}, s_{11}, t_{00}, t_{01}, t_{10}, t_{11}.$$

It maps to $\mathbb{R}^8$ with the coordinates:

$$p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}.$$

Algebraic representation of HMMs (4)

The map $f : \mathbb{R}^8 \to \mathbb{R}^8$ is given by the following:

$$\begin{aligned} f_{\sigma_1\sigma_2\sigma_3} = {} & s_{00}s_{00}t_{0\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{00}s_{01}t_{0\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{01}s_{10}t_{0\sigma_1}t_{1\sigma_2}t_{0\sigma_3} \\ & + s_{01}s_{11}t_{0\sigma_1}t_{1\sigma_2}t_{1\sigma_3} + s_{10}s_{00}t_{1\sigma_1}t_{0\sigma_2}t_{0\sigma_3} + s_{10}s_{01}t_{1\sigma_1}t_{0\sigma_2}t_{1\sigma_3} \\ & + s_{11}s_{10}t_{1\sigma_1}t_{1\sigma_2}t_{0\sigma_3} + s_{11}s_{11}t_{1\sigma_1}t_{1\sigma_2}t_{1\sigma_3}. \end{aligned}$$

The **tropicalization** of the map f is the map $g : \mathbb{R}^d \to \mathbb{R}^{\ell^n}$ defined by replacing products by sums and sums by minima in the formula of f. In the example, the tropicalization is the piecewise-linear map $g : \mathbb{R}^8 \to \mathbb{R}^8$:

$$\delta_{\sigma_1,\sigma_2,\sigma_3} = \min\{u_{h_1h_2} + u_{h_2h_3} + v_{h_1\sigma_1} + v_{h_2\sigma_2} + v_{h_3\sigma_3} : (h_1, h_2, h_3) \in \{0,1\}^3\}.$$

This minimum is attained by the most likely hidden data $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$, given the observations $(\sigma_1, \sigma_2, \sigma_3)$ and given the parameters $u_{..} = -\log(s_{..})$ and $v_{..} = -\log(t_{..})$. The sequence $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$ is known as **Viterbi sequence**.

Newton polytope & Minkowski sum

Let

$$f(\boldsymbol{x}) = \sum_k c_k \boldsymbol{x}^{\boldsymbol{a}_k}$$

where we use the shorthand notation $\boldsymbol{x}^{\boldsymbol{a}} = (x_1, .., x_n)^{(a_1,..,a_n)} = \prod_{i=1}^n x_i^{a_i}$. Then the Newton polytope associated to f is the convex hull of the $\{\boldsymbol{a}_k\}_k$ that is

$$\left\{ \sum_k \alpha_k \boldsymbol{a}_k : \sum_k \alpha_k = 1 \ \ \& \ \ \forall j \ \ \alpha_j \geq 0 \right\}.$$

The Minkowski sum of two sets of position vectors A and B in Euclidean space is formed by adding each vector in A to each vector in B:

$$A + B = \{a + b | a \in A, b \in B\}$$

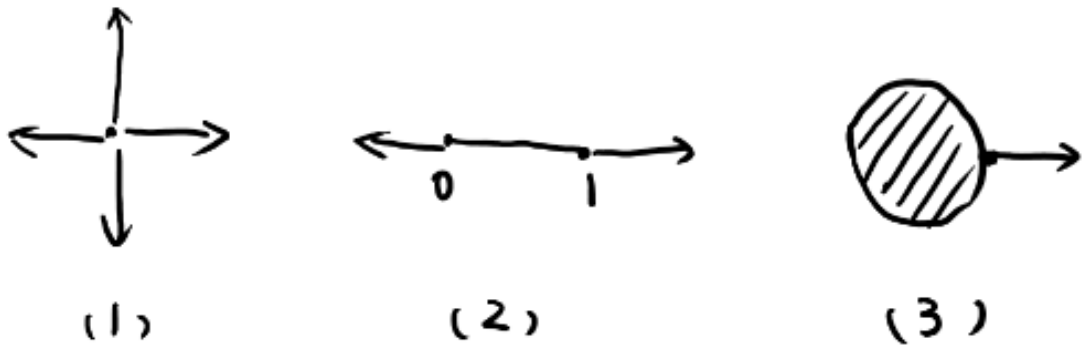
Normal cone - from Convex optimization (1)

Let $S \subset \mathbb{R}^n$ be a closed, convex set. The **normal cone** of S is the set-valued mapping $N_S : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, given by

$$N_S(x) = \begin{cases} \{g \in \mathbb{R}^n | (\forall z \in S) \;\; g^T(z - x) \leq 0\}, & \text{if } x \in S \\ \emptyset, & \text{if } x \notin S \end{cases}$$

Note: I was confused by the algebraic geometry definition, but I think convex definition here still works for this particular case of Newton polytope (convex hull) and more intuitive. For more, please check the book of Steven Boyd.

Normal cone - from Convex optimization (2)



1. *Let* $S = \{z\}$.

$$N_S(x) = \begin{cases} \mathbb{R}^n & if\ x = z \\ \emptyset & otherwise \end{cases}$$

2. *Let* $S = [0, 1]$.

$$N_S(x) = \begin{cases} \mathbb{R}_{\leq 0} & if\ x = 0 \\ \mathbb{R}_{\geq 0} & if\ x = 1 \\ \{0\} & if\ x \in (0, 1) \\ \emptyset & otherwise \end{cases}$$

3. *Let* $S = \{x \mid \|x\| \leq 1,\ x \in \mathbb{R}^n\}$.

$$N_S(x) = \begin{cases} \mathbb{R}_{\geq 0}x & if\ \|x\| = 1 \\ \{0\} & if\ \|x\| < 1 \\ \emptyset & otherwise \end{cases}$$

Normal fans - from Convex optimization (1)

For each vertex v of a polytope, we define the set E of its edges oriented towards its neighbors. With $E = \{e_1, .., e_n\}$ we build a polyhedral cone $C(v)$ named the primal cone of v:

$$C(v) = \{u_1 e_1 + .. + u_n e_n, \forall u_j \geq 0\}.$$

Indeed, a polytope can be written as the intersection of all the primal cones attached to its vertices:

$$A = \bigcup_{v_i \in \mathcal{V}_A} C(v_i)$$

where A is a polytope and $\mathcal{V}_A$ is the set of its vertices.

Normal fans - from Convex optimization (2)

For each vertex v of a polytope, we define the set $N = \{n_1, .., n_k\}$ of the outer normals of its corresponding facets. We build the **dual** (polyhedral) cone $C_D(v)$ at v:

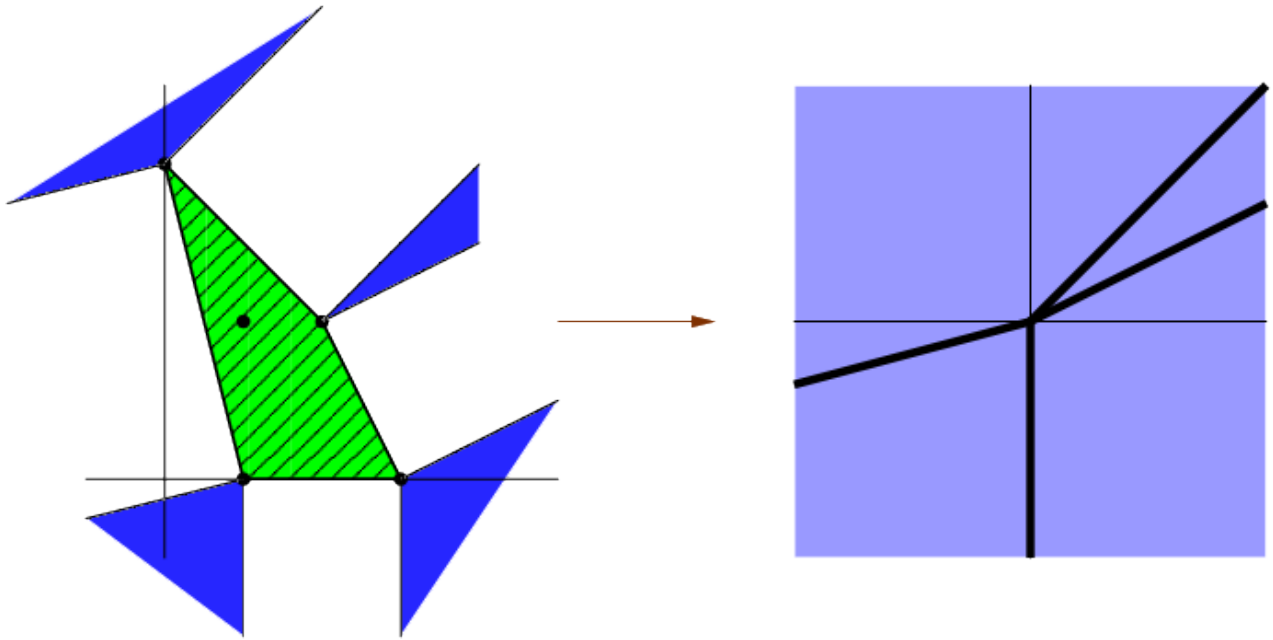
$$C_D(v) = \{t_1 n_1 + .. + t_k n_k, \forall t_i \geq 0\}.$$

The normal fan of a polytope A is the set of all the dual cones:

$$N(A) = \bigcup_{v \in \mathcal{V}_A} C_D(v),$$

that forms a partition of the whole space.

Normal fans - from Convex optimization (3)



Source: Elizalde Sergi and Woods Kevin, *Bounds on the number of inference functions of a graphical model*, Statistica Sinica 17 (2006).

Algebraic representation of HMMs (5) - The key!

The key observation, which we discuss in more detail in section 4, is that the set of parameters (U, V) that selects the Viterbi sequence $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$ is the normal cone at a vertex of the Newton polytope of the polynomial $f_{\sigma_1\sigma_2\sigma_3}$. This polytope is four-dimensional, it has eight vertices, and its normal fan represents the solution to problem 4 in section 1 when $\sigma = \sigma_1\sigma_2\sigma_3$ is fixed.

We can also consider an extension of problem 4 in which $\sigma = \sigma_1\sigma_2\sigma_3$ ranges over all possible observations. The solution is given by the Newton polytope of the map f. In our example, it is a five-dimensional polytope with 398 vertices, 1,136 edges, 1,150 two-faces, 478 three-faces, and 68 facets, namely, the Minkowski sum of eight copies of the earlier four-dimensional polytope for $(\sigma_1, \sigma_2, \sigma_3) \in \{0, 1\}^3$. For a concrete numerical example, fix the parameters $U^* = \left(\begin{smallmatrix} 6 & 5 \\ 8 & 1 \end{smallmatrix}\right)$ and $V^* = \left(\begin{smallmatrix} 0 & 8 \\ 8 & 8 \end{smallmatrix}\right)$. We find the following:

If the observed string at $Y_1Y_2Y_3$ is

$$\sigma_1\sigma_2\sigma_3 = 000\ 001\ 010\ 011\ 100\ 101\ 110\ 111,$$

then the Viterbi sequence at $X_1X_2X_3$ is

$$\hat{h}_1\hat{h}_2\hat{h}_3 = 000\ 001\ 000\ 011\ 000\ 111\ 110\ 111.$$

Complexity of Newton polytopes of graphical models (1)

Consider a directed and acyclic graphical model with E edges and n observed random variables $Y_1, .., Y_n$, each might take ℓ possible values. Such a model is given by a polynomial map $f : \mathbb{R}^d \to \mathbb{R}^{\ell^n}$. Each coordinate f_σ of f is a polynomial of degree at most E in the model parameters $s_1, .., s_d$.

Consider any of the ℓ^n possible observations σ. We have

$$f_\sigma(s_1, .., s_d) = P(\boldsymbol{Y} = \sigma)$$

is the probability of making this particular observation. Let $u_i = -\log(s_i)$. By **tropicalization**, we have:

$$g_\sigma(u_1, .., u_d) = -\log P(\boldsymbol{X} = \hat{h} | \boldsymbol{Y} = \sigma)$$

where $\hat{h}$ **maximizes** $P(\boldsymbol{X} = h | \boldsymbol{Y} = \sigma)$ (e.g., min of negatives turns into max).

Complexity of Newton polytopes of graphical models (2)

The domains of linearity of the functions g_σ are the cones in the normal fan of the Newton polytope of f_σ. Given observation σ, the **explanation** is the hidden values $\hat{h}$ that maximizes $P(\boldsymbol{X} = h | \boldsymbol{Y} = \sigma)$ for any of the parameters $(u_1, .., u_d)$.

Each logarithmic parameter vector $\boldsymbol{u} = (u_1, .., u_d)$ defines an **inference function** $\sigma \to \hat{h}$ from the set of observations to the set of explanations.

Complexity of Newton polytopes of graphical models (3)

For the HMM, each inference function $\{1,..,\ell\}^n \to \{1,..,k\}^n$ takes an observed sequence σ to the corresponding Viterbi sequence $\hat{h}$. There are $(k^n)^{\ell^n} = k^{n\ell^n}$ such functions, but most of them are **not** inference function.

This is a generous upper bound (without any knowledge of the graph topology). The main theme of this paper is to get a tighter bound.

Consider the binary HMM of length 3. There are $8^8 \approx 16 \times 10^6$ Boolean functions $\{0,1\}^3 \to \{0,1\}^3$, but only 398 are inference functions that is the number of vertices of the Newton polytope of the map f.

Complexity of Newton polytopes of graphical models (4)

Proposition 6. *The inference functions $\sigma \mapsto \hat{\mathbf{h}}$ of a graphical model f are in bijection with the vertices of the Newton polytope of the map f. The explanations $\hat{\mathbf{h}}$ for a fixed observation σ in a graphical model are in bijection with the vertices of the Newton polytope of the polynomial f_σ.*

Theorem 7. *Consider graphical models f whose number of parameters d is fixed and whose number n of observed random variables and number of edges E varies. (Typically, E is a linear function of n.) Then, the number of vertices of the Newton polytope $NP(f_\sigma)$ of f_σ is bounded above by the following:*

$$\begin{aligned}\text{No. of vertices } (NP(f_\sigma)) &\leq \text{constant} \cdot E^{d(d-1)/(d+1)} \\ &\leq \text{constant} \cdot E^{d-1}.\end{aligned}$$

Complexity of Newton polytopes of graphical models (5)

Proposition 6. *The inference functions $\sigma \mapsto \hat{\mathbf{h}}$ of a graphical model f are in bijection with the vertices of the Newton polytope of the map f. The explanations $\hat{\mathbf{h}}$ for a fixed observation σ in a graphical model are in bijection with the vertices of the Newton polytope of the polynomial f_σ.*

Theorem 7. *Consider graphical models f whose number of parameters d is fixed and whose number n of observed random variables and number of edges E varies. (Typically, E is a linear function of n.) Then, the number of vertices of the Newton polytope $NP(f_\sigma)$ of f_σ is bounded above by the following:*

$$\begin{aligned}\text{No. of vertices } (NP(f_\sigma)) &\leq \text{constant} \cdot E^{d(d-1)/(d+1)} \\ &\leq \text{constant} \cdot E^{d-1}.\end{aligned}$$

Complexity of Newton polytopes of graphical models (6)

Corollary 8. *For any fixed observation in the homogeneous HMM, the number of explanations is at most $C_{k,l} \cdot n^{k(k+l)}$. If all random variables are binary, then the upper bound $C \cdot n^{10/3}$ holds.*

The meaning is the number of explanations is a polynomial in terms of number of observed nodes (values).

Corollary 10. *The number of inference functions of a graphical model is at most $l^{nC_d} E^{d-1}$; hence, this number scales at most singly exponentially in the complexity (n, E) of the graphical model.*

The number of inference functions is **not** bounded by a polynomial (e.g., ℓ^n). The number of Boolean functions $\{0,1\}^n \to \{0,1\}^n$ is 2^{n2^n}, but the number of inference functions is at most $2^{\text{poly(n)}}$.

Complexity of Newton polytopes of graphical models (7)

In practical applications of graphical models, it may be infeasible to compute all (singly exponentially many) inference functions. Nonetheless, we believe that important insight can be gained by computing and classifying the Newton polytopes of graphical models f on few random variables. Such a study would be the polyhedral analog to the algebraic classification of ref. 1.

However, for a fixed observation σ, the size of the Newton polytope of f_σ grows polynomially with the size of the graphical model, and therefore, there is hope that the polytopes can be computed efficiently. Despite the fact that the Newton polytope of f_σ has polynomially many vertices in the size of the graphical model, the number of terms in f_σ grows exponentially. This is a potential problem because the computation of the Newton polytope requires the inspection of these terms. The following result states that the convex hull computations scale with the running time of the sum–product algorithm, which for many models of interest scales polynomially with the size of the graphical model.

Proposition 11 (Polytope Propagation). *The Newton polytopes of the polynomials f_σ can be computed recursively by using the decomposition of f_σ according to the sum–product algorithm.*

General Markov Model on Binary Tree (1)

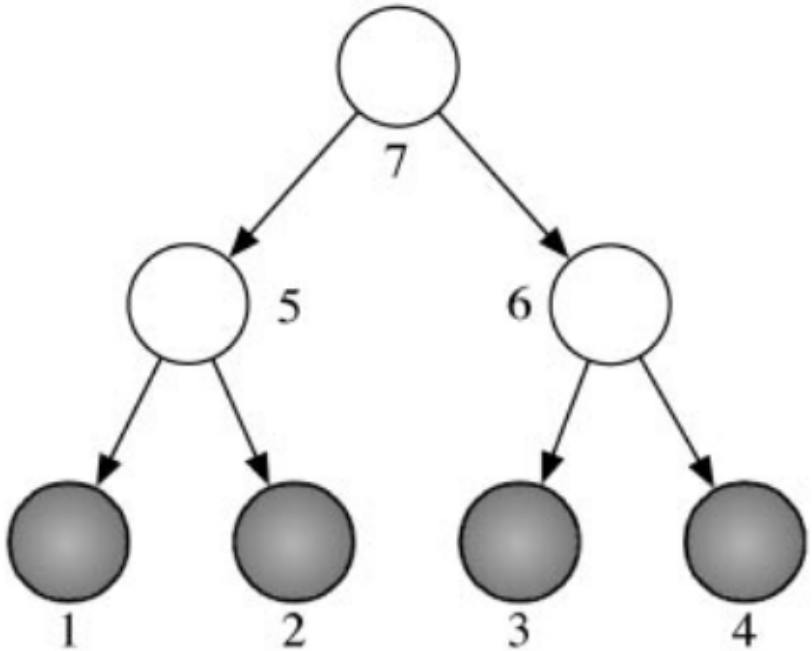


Fig. 3. A directed binary tree with $n = 4$ leaves.

HMM is indeed a binary-tree graphical model!

Directed tree τ with observed random variables $Y_1, .., Y_n$ at the leaves. Each edge e has a different transition matrix $S^e = [s^e_{\mu v}]$, an arbitrary distinct $\ell \times \ell$ matrix (the general model given by Allman and Rhodes).

General Markov Model on Binary Tree (2)

Proposition 12. *The general Markov model for the binary tree τ is the image of a map f: $\mathbf{R}^{(2n-2)l^2} \to \mathbf{R}^{l^n}$, where each coordinate of f is a multilinear polynomial in the unknowns $\{(s^e_{\mu\nu}),\ e \text{ edge of } \tau\}$.*

If we denote an edge between nodes i and j by (ij), and τ' is the tree τ without the leaves, then the coordinate of the multilinear map f indexed by an observed sequence $(\sigma_1, \ldots, \sigma_n)$ can be written as follows:

$$p_{\sigma_1 \cdots \sigma_n} = \sum_h \prod_{\substack{i \in \tau' \\ \text{with children } j,k}} (s^{(ij)}_{h_i h_j} \cdot s^{(ik)}_{h_i h_k}) \,. \qquad [5]$$

Tropicalization leads to the sum-product algorithm with ordinary arithmetic $(+, \times)$ replaced by tropical arithmetic $(\min, +)$:

$$\delta_{\sigma_1 \cdots \sigma_n} = \min_h \sum_{\substack{i \in \tau' \\ \text{with children } j,k}} (v^{(ij)}_{h_i h_j} + v^{(ik)}_{h_i h_k}). \qquad [9]$$

General Markov Model on Binary Tree (3)

Proposition 14. *The number of vertices of the Newton polytope of any coordinate f_σ in the homogeneous tree model is bounded above by n^{l^2-1} times a constant depending only on l.*

That means the number of inference functions (the number of vertices of the Newton polytope) is bounded by a polynomial of the number of observed variables: good!