Tropical Geometry of Statistical Models

Lior Pachter and Bernd Sturmfels
Department of Mathematics, University of California, Berkeley, CA 94720

February 4, 2008

Abstract

This paper presents a unified mathematical framework for inference in graphical models, building on the observation that graphical models are algebraic varieties. From this geometric viewpoint, observations generated from a model are coordinates of a point in the variety, and the sum-product algorithm is an efficient tool for evaluating specific coordinates. The question addressed here is how the solutions to various inference problems depend on the model parameters. The proposed answer is expressed in terms of tropical algebraic geometry. A key role is played by the Newton polytope of a statistical model. Our results are applied to the hidden Markov model and to the general Markov model on a binary tree.

1 Algebraic Statistics, Tropical Geometry, and Inference

This paper presents a unified mathematical framework for probabilistic inference with statistical models, such as graphical models. Our approach is summarized as follows:

(a) Statistical models are algebraic varieties.

(b) Every algebraic variety can be tropicalized.

(c) Tropicalized statistical models are fundamental for parametric inference.

By a *statistical model* we mean a family of joint probability distributions for a collection of discrete random variables $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$. Thesis (a) states that many families of interest can be characterized by polynomials in the joint probabilities $p_{\sigma_1 \cdots \sigma_n} = \text{Prob}(Y_1 = \sigma_1, \ldots, Y_n = \sigma_n)$. The emerging field of algebraic statistics [12, 19] offers algorithms for this polynomial representation.

Tropicalization means replacing the arithmetic operations $(+, \times)$ by the operations $(\min, +)$. This process captures the essence of what happens when the joint probabilities $p_{\sigma_1 \cdots \sigma_n}$ are replaced by their logarithms. The tropicalization of an algebraic variety is a piecewise-linear set which enjoys many features familiar from algebraic geometry [8, 17]. In particular, the tropicalization of a statistical model is a piecewise-linear set in the space with logarithmic coordinates $-\log(p_{\sigma_1 \cdots \sigma_n})$.

Thesis (c) states that tropical algebraic geometry of statistical models is fundamental in analyzing the behavior of inference algorithms under the variation of model parameters. By *inference* we mean the evaluation of one or more coordinates of a single point on the algebraic variety, in either $(+, \times)$ or $(\min, +)$ arithmetic. This is the standard notion of inference used for graphical models in statistical learning theory [15], but it differs from other (more classical) notions of inference in mathematical statistics. By *parametric inference* we mean the analysis of the dependence of inference on parameters.

To give a more concrete discussion of parametric inference it is useful to focus on directed graphical models. A *directed graphical model* (or *Bayesian network*) is a finite directed acyclic graph G with two kinds of vertices, *observed variables* $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$ and *hidden variables* $\mathbf{X} = \{X_1, \ldots, X_m\}$, where each

edge is labeled by a transition matrix whose entries are linear forms in some parameters. The rules of discrete probability express the observed probabilities $p_{\sigma_1 \cdots \sigma_n}$ as polynomials of degree E in the parameters, where E is the number of edges of G. The polynomials parametrize the graphical model as an algebraic variety.

The two standard types of inference questions for graphical models are:

1. the calculation of *marginal probabilities*:

$$p_{\sigma_1 \cdots \sigma_n} \quad = \quad \sum_{h_1,\ldots,h_m} \text{Prob}(X_1 = h_1, \ldots, X_m = h_m, Y_1 = \sigma_1, \ldots, Y_n = \sigma_n),$$

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

$$\delta_{\sigma_1 \cdots \sigma_n} \quad = \quad \min_{h_1,\ldots,h_m} -\log\left(\text{Prob}(X_1 = h_1, \ldots, X_m = h_m, Y_1 = \sigma_1, \ldots, Y_n = \sigma_n)\right),$$

where the h_i range over all the possible assignments for the hidden random variables X_i. Together, these two primitives can be used to effectively solve a range of other inference problems, including the calculation of conditional probabilities and other quantities of interest. The key to inference in graphical models is the *sum-product algorithm* [14] (also known as the *generalized distributive law* [3]). This polynomial-time algorithm is used, both in ordinary arithmetic $(+,\times)$ and in tropical arithmetic $(\min,+)$, to *efficiently* solve Problems 1 and 2. For more background on the sum-product algorithm, and for connections to message passing and the junction tree algorithm see [15].

Although the sum-product algorithm provides efficient solutions to the basic inference problems 1 and 2, it only applies to one coordinate $p_{\sigma_1 \cdots \sigma_n}$ of one distribution at a time. What we are interested in are the *parametric* versions of the inference problems. They can be phrased as follows:

3. Find all parameters for a model which result in the same values for all $p_{\sigma_1,\cdots,\sigma_n}$.

4. Given observations $\mathbf{Y} = \sigma$ and hidden data $\mathbf{X} = \mathbf{h}$, identify all parameters such that $\mathbf{h}$ is the most likely explanation for the observations σ.

As we will see, the following *modeling* questions are fundamentally related to Problems 3 and 4:

5. Which (parameter independent) relations on the probabilities $p_{\sigma_1 \cdots \sigma_n}$ does the model imply?

6. Describe the tropicalization of the variety corresponding to a graphical model.

Problem 5 asks for the ideal of *polynomial invariants* of a statistical model [12]. Invariants have been investigated in phylogenetics [2, 5] where they can help to identify good trees for aligned DNA sequences.

The primary goal of our study is to give a practical answer to question 4 for graphical models. Our main algorithmic result is an efficient procedure for parametric inference that can be viewed as a polytopal analog of the sum-product algorithm. The efficiency is based on the complexity estimates for Newton polytopes which we derive in Section 4. The resulting *polytope propagation algorithm* is applied to problems in biological sequence analysis in the companion paper [18].

The mathematics to be developed in Sections 3 and 4 is of independent interest. It also furnishes new tools for parametric inference (Problems 3 and 4) and parametric modeling (Problems 5 and 6) which are applicable to a wide range of statistical problems. We demonstrate this by analyzing the hidden Markov model (HMM) and the general Markov model on a binary tree, in Sections 2 and 5 respectively.

2 Algebraic Representation of Hidden Markov Models

A graphical model is an algebraic variety which is presented as the image of a highly structured polynomial map $f : \mathbf{R}^d \to \mathbf{R}^m$. Here $\mathbf{R}^d$ is the space whose coordinates are the model parameters $s_1, \ldots, s_d$ and $\mathbf{R}^m$ is the space whose coordinates $p_\sigma = p_{\sigma_1 \cdots \sigma_n}$ are the joint probabilities for the observed random variables. In applications, the integer m is much larger than the integer d, in fact; it is so large that one can only look at one coordinate p_σ at a time. Each coordinate $f_\sigma = f_\sigma(s_1, \ldots, s_d)$ of the map f is a polynomial function in $s_1, \ldots, s_d$. The efficient evaluation of these functions relies on the sum-product algorithm. Here we study the (parametric) inference and modeling problems in the familiar context of the *hidden Markov model* (HMM).

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

$$\begin{aligned} p(X_i \,|\, X_1, X_2, \ldots, X_{i-1}) &= p(X_i \,|\, X_{i-1}), \\ p(Y_i \,|\, X_1, \ldots, X_i, Y_1, \ldots, Y_{i-1}) &= p(Y_i \,|\, X_i). \end{aligned}$$

We consider the homogeneous model with uniform initial distribution, where all transitions $X_i \to X_{i+1}$ are given by the same $k \times k$-matrix $S = (s_{ij})$ and all transitions $X_i \to Y_i$ are given by the same $k \times l$-matrix $T = (t_{ij})$. Throughout our discussion we disregard for simplicity the usual probabilistic hypothesis that S and T are non-negative and all row sums are 1.

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

Problem 3 is to compute the fibers of the map f. In statistics, this is called *parameter identification*. We use the term *coordinate polynomials* for the polynomials f_σ that are coordinates of the map f.

Our running example in this section is the case $n = 3$ with binary random variables $(k = l = 2)$. The graph of this model is drawn in Figure 1. The shaded nodes are the observed random variables.

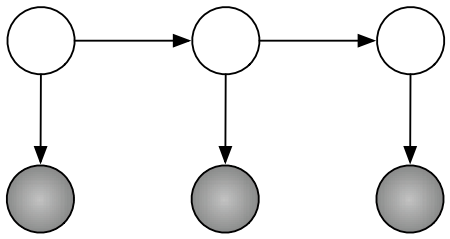

Figure 1: The hidden Markov model of length three.

Here the parameter space is $\mathbf{R}^8$ with coordinates $s_{00}, s_{01}, s_{10}, s_{11}, t_{00}, t_{01}, t_{10}, t_{11}$, and it maps to $\mathbf{R}^8$ with coordinates $p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111}$. The map $f : \mathbf{R}^8 \to \mathbf{R}^8$ is given by

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

The hidden Markov model (i.e. the image of f) is the zero set of the quartic polynomial

$$\begin{aligned}
&p_{011}^2 p_{100}^2 - p_{001}^2 p_{110}^2 + p_{000} p_{011} p_{101}^2 - p_{000} p_{101}^2 p_{110} + p_{000} p_{011} p_{110}^2 - p_{001} p_{010}^2 p_{111} + p_{001}^2 p_{100} p_{111} \\
&\quad + p_{010}^2 p_{100} p_{111} - p_{001} p_{100}^2 p_{111} - p_{000} p_{011}^2 p_{110} - p_{001} p_{011} p_{100} p_{101} - p_{010} p_{011} p_{100} p_{101} \\
&+ p_{001} p_{010} p_{011} p_{110} - p_{010} p_{011} p_{100} p_{110} + p_{001} p_{010} p_{101} p_{110} + p_{001} p_{100} p_{101} p_{110} + p_{000} p_{010} p_{011} p_{111} \\
&- p_{000} p_{011} p_{100} p_{111} - p_{000} p_{001} p_{101} p_{111} + p_{000} p_{100} p_{101} p_{111} + p_{000} p_{001} p_{110} p_{111} - p_{000} p_{010} p_{110} p_{111}.
\end{aligned}$$

This polynomial was found by a *Gröbner basis* computation. See the discussion on *implicitization* in [7, §3].

In general, the polynomial functions on $\mathbf{R}^{l^n}$ which vanish on the image of f are the called *invariants of the model*. They form a prime ideal I_f. In our example, I_f is generated by the quartic polynomial above. Problem 5 is to compute generators of the ideal I_f. When l^n and d are small, this can be done using Gröbner bases, and in some cases it is possible to characterize I_f based on the structure of the model (see, for example, Conjecture 13), but in general Problem 5 is hard and the ideal I_f may remain unknown.

Here is where tropical geometry comes in. The *tropicalization* of our map f is the map $g : \mathbf{R}^d \to \mathbf{R}^{l^n}$ defined by replacing products by sums and sums by minima in the formula for f. In our example ($n = 3, k = l = 2$), the tropicalization is the piecewise-linear map $g : \mathbf{R}^8 \to \mathbf{R}^8$, $(U,V) \mapsto \delta$ with

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

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 $u_{\cdot\cdot} = -\log(s_{\cdot\cdot})$ and $v_{\cdot\cdot} = -\log(t_{\cdot\cdot})$. The sequence $(\hat{h}_1, \hat{h}_2, \hat{h}_3)$ is known as the *Viterbi sequence* in the HMM literature [20]. It solves Problem 2 in the Introduction.

The key observation, which we discuss in more detail in Section 4, is that the set of parameters (U,V) which select 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 4-dimensional, it has 8 vertices, and its normal fan represents the solution to Problem 4 in the Introduction when $\sigma = \sigma_1\sigma_2\sigma_3$ is fixed.

We can also consider an extension of Problem 4 where $\sigma = \sigma_1\sigma_2\sigma_3$ ranges over all possible observations. The solution is given by the Newton polytope of the map f. In our example, this is a 5-dimensional polytope with 398 vertices, 1136 edges, 1150 two-faces, 478 three-faces and 68 facets, namely, the Minkowski sum of eight copies of the earlier 4-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:

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

The set of all parameters (U,V) leading to the same conclusions as (U^*,V^*) is the cone defined by

$$\begin{aligned}
&u_{01} - u_{00} + v_{11} - v_{01} \le 0,\ u_{10} - u_{11} + v_{00} - v_{10} \le 0,\ u_{00} + v_{01} - u_{10} - v_{11} \le 0, \\
&2u_{00} + v_{01} - u_{01} - u_{10} - v_{11} \le 0,\ 2u_{11} + v_{10} + v_{11} - u_{00} - u_{01} - v_{00} - v_{01} \le 0.
\end{aligned}$$

Our solution to the parametric inference problem with respect to all observations simultaneously consists of 398 such cones. The *tropical HMM* is the union of the images of these cones under the piecewise-linear map $g : (U,V) \mapsto \delta$. This image is a piecewise-linear set of dimension 7. The cone which contains the chosen parameters (U^*,V^*) mapped to a 7-dimensional cone in the tropical HMM (it spans the hyperplane $\delta_{010} = \delta_{100}$) but most of the other 397 cones are mapped to lower-dimensional cones by the map g. The question how the number 398 grows as the length n increases will be addressed in Corollary 10.

3 Positivity and Morphisms in Tropical Geometry

We have seen that a graphical model is the image of a polynomial map f from the space of parameters to the space of joint probability distributions on the observed random variables. Furthermore, we have seen that the tropicalization of f arises naturally in solving Problem 4. In this section we study the geometry of tropicalization in the more general setting where $f : \mathbf{R}^d \to \mathbf{R}^m$ is an arbitrary polynomial map. In statistical applications, it is usually the case that each coordinate f_σ of the map f is a polynomial with positive coefficients. If this holds then the polynomial map f is called *positive*. We say that f is *surjectively positive* if, in addition, f maps the positive orthant surjectively onto the positive points in the image, in symbols,

$$f(\mathbf{R}^d_{>0}) \quad = \quad \text{image}(f) \cap \mathbf{R}^m_{>0}. \tag{2}$$

The set of all polynomial functions which vanish on the image of f is a prime ideal I_f in the polynomial ring $\mathbf{R}[p_1,\ldots,p_m]$. The closure of the image of f is the variety of the prime ideal I_f.

In tropical geometry, we replace the variety of I_f by a piecewise-linear set as follows. The *tropical variety* $\mathcal{T}(I_f)$ is the set of all weight vectors $w \in \mathbf{R}^m$ such that the initial ideal $\text{in}_w(I_f)$ contains no monomial [17, 22]. Following [21], we define the *positive tropical variety* $\mathcal{T}^+(I_f)$ as the set of all weight vectors $w \in \mathbf{R}^m$ such that the initial ideal $\text{in}_w(I_f)$ contains no polynomial with only positive coefficients. The tropical variety $\mathcal{T}(I_f)$ is a *polyhedral fan* in $\mathbf{R}^m$, and $\mathcal{T}^+(I_f)$ is a *polyhedral subcomplex* of $\mathcal{T}(I_f)$. This means that $\mathcal{T}(I_f)$ is a finite union of closed convex polyhedral cones that fit together nicely, and $\mathcal{T}^+(I_f)$ is the union of a subset of these cones. The *tropicalization* of the polynomial map f is the piecewise-linear map $g : \mathbf{R}^d \to \mathbf{R}^m$ defined by replacing products by sums and sums by minima in the evaluation of f. We say that g is a *tropical morphism*. Examples of tropical morphisms appear in the displayed formulas (1), (3), (4), (5), (10) and (11).

The following theorem describes the geometry of this situation. We define the *Newton polytope* of a polynomial map $f : \mathbf{R}^d \to \mathbf{R}^m$ as the Minkowski sum in $\mathbf{R}^d$ of the Newton polytopes of its coordinates $f_1,\ldots,f_m$. For basics on Newton polytopes and their normal fans see [22, §1].

Theorem 2. *The tropical morphism g is linear on each cone in the normal fan of the Newton polytope of f. Its image is a fan contained in $\mathcal{T}(I_f)$. If f is positive then* $\text{image}(g)$ *is a subset of $\mathcal{T}^+(I_f)$, but it is generally not a polyhedral subcomplex. If f is surjectively positive then* $\text{image}(g) = \mathcal{T}^+(I_f)$.

Proof. Let P_i denote the Newton polytope of the polynomial $f_i = f_i(s_1,\ldots,s_d)$. By definition, P_i is the convex hull in $\mathbf{R}^d$ of all non-negative lattice points $a = (a_1,\ldots,a_d) \in \mathbf{N}^d$ such that the monomial $s_1^{a_1}\cdots s_d^{a_d}$ appears with non-zero coefficient in f_i. The piecewise-linear concave function g_i is the *support function* of the polytope P_i. This means that $g_i(w)$ is the minimum value attained on P_i by the linear functional $a \mapsto w \cdot a$. In particular, the function $g_i : \mathbf{R}^d \to \mathbf{R}$ is linear on each cone in the normal fan of P_i.

The Newton polytope of the map f is the Minkowski sum $P_1 + \cdots + P_m = \{a_1 + \cdots + a_m : a_i \in P_i\}$. The normal fan of $P_1 + \cdots + P_m$ is the common refinement of the normal fans of $P_1,\ldots,P_m$. This shows that the function $f = (f_1,\ldots,f_m) : \mathbf{R}^d \to \mathbf{R}^d$ is linear on each cone of the normal fan of the Newton polytope of f. Since g is continuous, the image of g is a closed polyhedral fan in $\mathbf{R}^m$.

Consider any vector $w \in \mathbf{R}^d$. We must show that $g(w)$ lies in $\mathcal{T}(I_f)$, and if f is positive then $g(w)$ lies in $\mathcal{T}^+(I_f)$. Let ϕ be any polynomial in the ideal I_f. If we substitute $p_1 = f_1,\ldots,p_m = f_m$ into $\phi = \phi(p_1,\ldots,p_m)$ then we get zero. Consequently, if we substitute the initial forms $p_1 = \text{in}_w(f_1),\ldots,p_m = \text{in}_w(f_m)$ into the initial form $\text{in}_{g(w)}(\phi)$ then the result is zero. See equation (11.2) on page 100 in [22]. This implies that $\text{in}_{g(w)}(\phi)$ is not a monomial. Moreover, if f is positive then ϕ must have two terms whose coefficients have opposite signs. This implies the desired conclusion.

The following example shows that image(g) need not be a subcomplex of $\mathcal{T}^+(I_f)$. If f is assumed to be surjectively positive, then it follows from [21, Proposition 2.5] that image$(g) = \mathcal{T}^+(I_f)$. □

Example 3. Let $d = 3$, $m = 4$ and consider the linear map

$$f : \mathbf{R}^3 \to \mathbf{R}^4,\ (s_1,s_2,s_3) \mapsto \big(s_1+s_2+s_3,\, s_1+2s_2+s_3,\, s_2+s_3,\, s_3\big).$$

Then I_f is the principal ideal generated by the linear form $p_1 - p_2 + p_3 - p_4$, and $\mathcal{T}(I_f)$ is essentially the normal fan of a tetrahedron. We identify $\mathcal{T}(I_f)$ with the complete graph K_4. The six edges of K_4 are labeled with six monomial-free initial ideals of I_f, namely,

$$\langle p_1+p_3\rangle, \langle -p_2-p_4\rangle, \langle p_1-p_2\rangle, \langle p_1-p_4\rangle, \langle -p_2+p_3\rangle, \langle p_3-p_4\rangle.$$

The first two of these six initial ideals contain a polynomial with positive coefficients. Hence the positive tropical variety $\mathcal{T}^+(I_f)$ is the four-cycle in K_4 formed by the remaining four edges.

The tropicalization of the linear map f is the tropical morphism

$$g : \mathbf{R}^3 \to \mathbf{R}^4,\ (u_1,u_2,u_3) \ \mapsto\ \big(\min(u_1,u_2,u_3), \min(u_1,u_2,u_3), \min(u_2,u_3), u_3\big). \tag{3}$$

The image of g is the set of all vectors (a,a,b,c) with $a \le b \le c$. Each vector (a,a,b,c) with $a < b < c$ has the initial ideal $\langle p_1 - p_2\rangle$, so it lies on a particular edge of K_4. But the same edge also accounts for all vectors (a,a,b,c) with $a < c < b$, none of which is in the image of g. Thus image(g) is a closed segment which covers only half of the edge of K_4 indexed by $\langle p_1 - p_2\rangle$.

Here it is easy to replace f by a parameterization f' which is surjectively positive, for instance,

$$f' : \mathbf{R}^4 \to \mathbf{R}^4,\ (s_1,s_2,s_3,s_4) \mapsto \big(s_1+s_3, s_1+s_4, s_2+s_4,\, s_2+s_3\big).$$

$$g' : \mathbf{R}^4 \to \mathbf{R}^4, (u_1,u_2,u_3,u_4) \ \mapsto\ \big(\min(u_1,u_3), \min(u_1,u_4), \min(u_2,u_4), \min(u_2,u_3)\big). \tag{4}$$

We have $I_f = I_{f'}$ but now the tropical morphism g' maps onto the entire four-cycle $\mathcal{T}^+(I_f)$. □

In the rest of this section we examine Theorem 2 for a small but important graphical model, namely, the *naive Bayes model with two features* [12, §7]. There are two observed random variables Y_1 and Y_2 dependent on one hidden binary random variable X. The two observed variables take k and l possible values respectively. The parameterization f of this model is the map $f : \mathbf{R}^{2(k+l)} \mapsto \mathbf{R}^{kl}$ given by

$$p_{ij} \ = \ s_{i0}t_{0j} + s_{i1}t_{1j}.$$

Thus the model consists of all $k \times l$-matrices $P = (p_{ij})$ of the form $P = S \cdot T$ where S is a $k \times 2$-matrix and T is a $2 \times l$-matrix, i.e., the model consists of precisely the $k \times l$-matrices of rank ≤ 2.

Proposition 4. *The parameterization f of the naive Bayes model with two features is surjectively positive. The ideal I_f is generated by the 3×3-subdeterminants of the $k \times l$-matrix $P = (p_{ij})$.*

Proof. The map f being positive means that if P is any positive matrix of rank 2 then S and T can be chosen to be positive. This is a known result in linear algebra (see e.g. [6]). The same statement is false for rank ≥ 3, i.e., the parameterization of the naive Bayes model with three or more features is not surjectively positive. A well-known result in commutative algebra states that the $(r+1) \times (r+1)$-minors of a $k \times l$-matrix generate a prime ideal. The variety of this ideal is the set of $k \times l$-matrices of rank $\le r$. This our ideal I_f for $r = 2$. □

The objects of Theorem 2 have been studied in [8] and [9]. The tropical variety $\mathcal{T}(I_f)$ is the set of $k \times l$-matrices of *tropical rank* ≤ 2, and the tropical variety $\mathcal{T}^+(I_f) = \text{image}(g)$ is the set of $k \times l$-matrices of *Barvinok rank* ≤ 2. Develin [9] determines the combinatorics and topology of these spaces when $\min(k,l) = 3$. He shows that $\mathcal{T}(I_f)$ is shellable but $\mathcal{T}^+(I_f)$ can have torsion in its integral homology groups.

The Newton polytope of the map f is an interesting combinatorial object, namely, it is the $(kl-k-l+2)$-dimensional zonotope associated with the complete bipartite graph $K_{k,l}$. The Newton polytope of each coordinate f_{ij} is a line segment, and the zonotope is their Minkowski sum. The normal fan is the hyperplane arrangement $\{u_{i0}-u_{i1} = v_{1j}-v_{0j}\}$. Its maximal cones correspond to the acyclic orientations of the complete bipartite graph $K_{k,l}$. West [25] showed that the number of facets of such a cone can be any integer between $k+l-1$ and kl. The total number of cones equals $\sum_{i=1}^{k} S(k,i)(-1)^{l+i} i!(i+1)^l$, where $S(k,i)$ is the Stirling number of the second kind. Here, the tropical morphism g is given by

$$g_{ij} \;=\; \min\bigl(u_{i0}+v_{0j}, u_{i1}+v_{1j}\bigr). \tag{5}$$

The map $g : \mathbf{R}^{2(k+l)} \mapsto \mathbf{R}^{kl}$ is piecewise-linear with respect to the hyperplane arrangement. Recent work of Federico Ardilla (in preparation) gives a complete classification of all fibers of g.

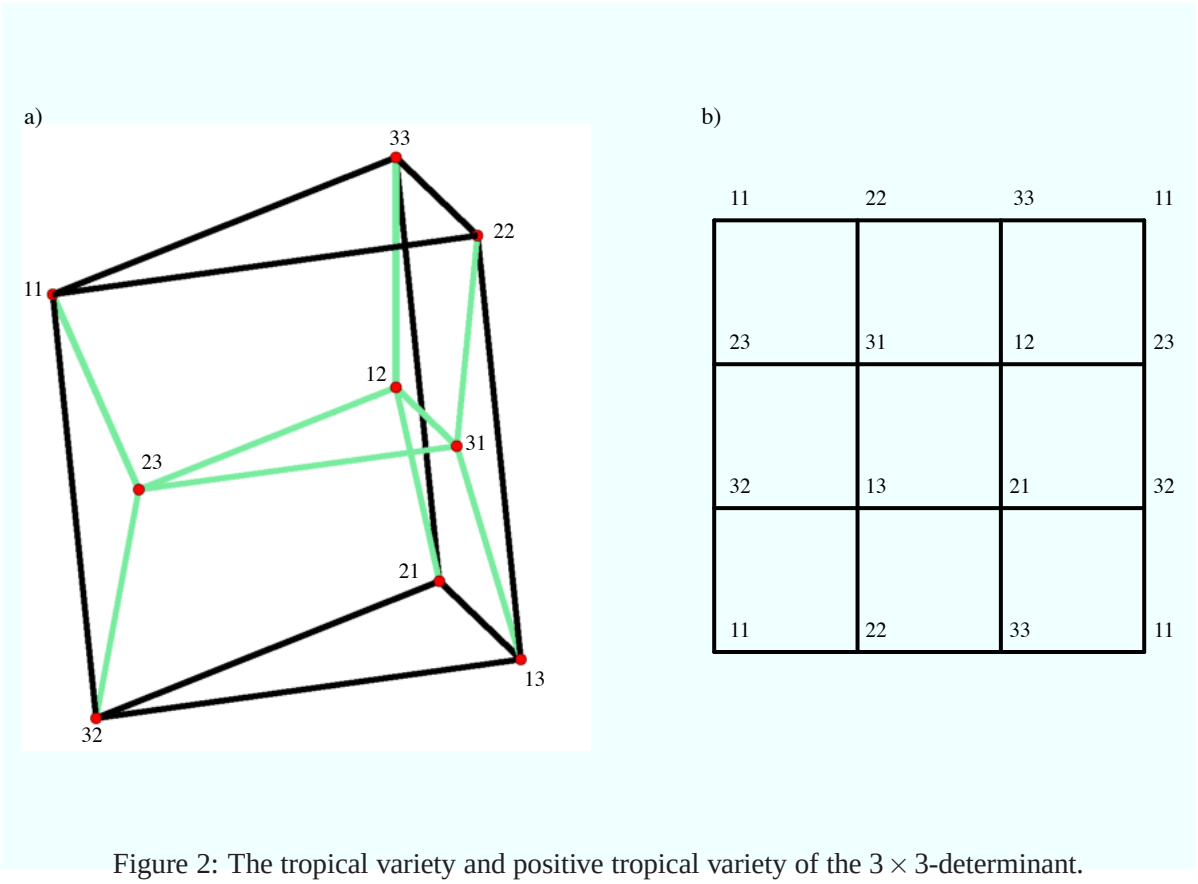

Figure 2: The tropical variety and positive tropical variety of the 3×3-determinant.

Example 5. Let $k = l = 3$, so the two observed random variables are ternary. The prime ideal is

$$I_f \;\;=\;\; \langle p_{11}p_{22}p_{33} - p_{11}p_{23}p_{32} - p_{12}p_{21}p_{33} + p_{12}p_{23}p_{31} + p_{13}p_{21}p_{32} - p_{13}p_{22}p_{31} \rangle.$$

The tropical variety $\mathcal{T}(I_f)$ is the fan over a two-dimensional polyhedral complex consisting of six triangles and nine quadrangles. This complex is the 2-skeleton of the product of two triangles, labeled as in Figure 2a. This complex is shellable. The positive tropical variety $\mathcal{T}^+(I_f)$ is the subcomplex consisting of the nine quadrangles shown in Figure 2b. Note that $\mathcal{T}^+(I_f)$ is a torus.

The Newton polytope of f is a five-dimensional zonotope with 230 vertices, one for each acyclic orientation of the complete bipartite graph $K_{3,3}$. The map g is linear on each of the 230 cones in the corresponding hyperplane arrangement, but it is rank-deficient on 68 of the cones. The remaining $162 = 18 \times 9$ cones are mapped onto the 9 quadrangles of the torus $\mathcal{T}^+(I_f)$. Thus the general fiber of g involves 18 cones. Of these, eight cones have 5 facets, eight cones have 6 facets, and two cones have 9 facets. □

4 Newton Polytopes of Graphical Models and their Complexity

Consider a graphical model with E edges and n observed random variables $Y_1, \ldots, Y_n$ each taking l values. Such a model is given by a positive polynomial map $f : \mathbf{R}^d \to \mathbf{R}^{l^n}$. Each coordinate f_σ of f is a polynomial of degree e in the model parameters $s_1, \ldots, s_d$. In this section we discuss the statistical meaning and the computational complexity of the mathematical objects introduced in the previous section.

We write $u_i = -\log(s_i)$ for the negative logarithms of the model parameters. Consider any of the l^n possible observations σ. The quantity $f_\sigma(s_1, \ldots, s_d)$ is the probability of making this particular observation, i.e. it is $\text{Prob}(\mathbf{Y} = \sigma)$. The quantity $g_\sigma(u_1, \ldots, u_d)$ is the negative logarithm of the conditional probability $\text{Prob}(\mathbf{X} = \widehat{\mathbf{h}} \,|\, \mathbf{Y} = \sigma)$ where $\widehat{\mathbf{h}}$ maximizes $\text{Prob}(\mathbf{X} = \mathbf{h} \,|\, \mathbf{Y} = \sigma)$ for the parameters $(s_1, \ldots, s_d)$. Clearly, the function $g_\sigma : \mathbf{R}^d \to \mathbf{R}$ is piecewise-linear and concave on the logarithmic parameter space.

The domains of linearity of the function g_σ are the cones in the normal fan of the Newton polytope of f_σ. Each maximal cone $\mathcal{C}$ is indexed by the hidden data $\widehat{\mathbf{h}}$ that maximizes $\text{Prob}(\mathbf{X} = \mathbf{h} | \mathbf{Y} = \sigma)$ for any of the parameters $(u_1, \ldots, u_d) \in \mathcal{C}$. The hidden data $\widehat{\mathbf{h}}$ which arise in this manner, for some choice of logarithmic parameters u, are called the possible *explanations* of the observation σ. For instance, for the hidden Markov model of Section 2, the explanations are the Viterbi sequences.

Let us now vary the observations. Each logarithmic parameter vector $\mathbf{u}$ defines an *inference function* $\sigma \mapsto \widehat{\mathbf{h}}$ from the set of observations to the set of explanations. For the HMM, each inference function $\{1, \ldots, l\}^n \to \{1, \ldots, k\}^n$ takes an observed sequence σ to the corresponding Viterbi sequence $\widehat{\mathbf{h}}$. There are $(k^n)^{l^n} = k^{n l^n}$ such functions, but most of these are **not** inference functions. For instance, consider the binary HMM of length three. There are $8^8 = 16,777,216$ Boolean functions $\{0,1\}^3 \to \{0,1\}^3$, but, as we have seen at the end of Section 2, only 398 of these are inference functions for the HMM.

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

In applications of graphical models, the number d of parameters and the number l of values of the observed random variables is small and fixed, but the number n of observed random variables is large. Recall that the model is the image of the map $f : \mathbf{R}^d \to \mathbf{R}^{l^n}$. Hence the dimension of the model remains fixed but the dimension of its ambient space grows exponentially in n. It is therefore algorithmically infeasible to compute the full tropical variety $\mathcal{T}(I_f)$. What we can do efficiently, however, is to compute the Newton polytopes of the f_σ, or even the Newton polytope of f. This allows us to glean information about the tropical variety from the domains of linearity of its "coordinate functions" g_σ.

Our next goal is to derive an upper bound on the number of vertices of the Newton polytopes.

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*

$$\#\text{vertices}(NP(f_\sigma)) \;\leq\; \text{constant} \cdot E^{d(d-1)/(d+1)} \;\leq\; \text{constant} \cdot E^{d-1}.$$

For many important families of graphical models, the number E of edges is bounded by a linear function in terms of the number n of observed nodes, and in those cases we can replace E by n. Hence, for any given observation σ, the number of explanations grows polynomially in n. For instance, in the hidden Markov model of Section 2 we have $E = 2n-1$, and a similar relationship holds in the tree model of Section 5.

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 proof of Theorem 7 and Corollary 8 are derived from the following classical result on lattice polytopes due to Andrews [1]. The necessary observation is that the Newton polytope of f_σ is contained in the cube $[0,E]^d$, and the volume of this cube equals E^d.

Proposition 9. (Andrews [1]) *For every fixed integer d there exists a constant C_d such that the number of vertices of any lattice polytope P in $\mathbf{R}^d$ is bounded above by $C_d \cdot \text{volume}(P)^{(d-1)/(d+1)}$.*

The Newton polytope of the map f was defined as the Minkowski sum of the l^n smaller Newton polytopes in Theorem 7. From this we infer the following naive bound on its number of vertices.

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.*

Consider the homogeneous HMM on binary random variables. Each inference function is a Boolean function $\{0,1\}^n \to \{0,1\}^n$, but not conversely. The number of all Boolean functions is 2^{n2^n}, which grows doubly exponentially in n. However, the number of inference functions is at most $2^{\text{polynomial}(n)}$.

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 analogue to the algebraic classification of [12].

On the other hand, 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 inspecting these terms. The following result states that, in fact, the convex hull computations scales 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 using the decomposition of f_σ according to the sum-product algorithm.*

Taken together, Theorem 7 and Proposition 11 say that **polytope propagation is an efficient algorithm for parametric inference with graphical models**. This statement is thesis (c) in our companion paper [18]. In that paper, the sum-product algorithm and the polytope propagation algorithm are explained and analyzed in more detail. We also demonstrate the practicality of our mathematical theory by explicitly computing (and statistically interpreting) various high-dimensional Newton polytopes for graphical models that arise in biological sequence analysis.

5 The General Markov Model on a Binary Tree

We conclude by illustrating the concepts we have developed in the context of tree Markov models. These are directed graphical models where the graph is a directed tree τ with observed random variables $Y_1, \ldots, Y_n$ at the leaves. The naive Bayes model in Section 3 is the special case where $n = 2$. Each edge e has a different transition matrix $S^e = [s^e_{\mu\nu}]$. We consider the general model in Allman and Rhodes [2], which means that the S^e are arbitrary distinct $l \times l$-matrices. In most applications, the transition matrices are from a special model family (e.g. in phylogenetics these may be Jukes-Cantor model or the Hasegawa-Kishino-Yano model). As before, we relax the hypothesis that transition probabilities are non-negative and sum to 1. Hence the $s^e_{\mu\nu}$ are distinct unknowns. For simplicity we shall further assume that the tree τ is binary.

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:

$$p_{\sigma_1 \cdots \sigma_n} \quad = \quad \sum_{h} \prod_{\substack{i \in \tau' \\ \text{with children } j,k}} \left(s^{(ij)}_{h_i h_j} \cdot s^{(ik)}_{h_i h_k} \right). \tag{6}$$

Here h ranges over all colorations $h = (h_i)_{i \in \tau}$ of the nodes such that $h_j = \sigma_j$ for all leaves j. Our running example in this section is the binary tree in Figure 3 with binary random variables $(l = 2)$.

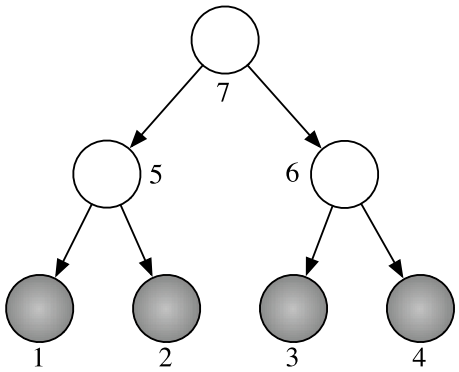

Figure 3: A directed binary tree with $n = 4$ leaves.

In this example, the coordinates of the multilinear map $f : \mathbf{R}^{24} \to \mathbf{R}^{16}$ are given by the formula

$$p_{\sigma_1\sigma_2\sigma_3\sigma_4} \quad = \quad \sum_{\{h_5,h_6,h_7\} \in \{0,1\}^3} (s^{(75)}_{h_7 h_5} \cdot s^{(76)}_{h_7 h_6}) \cdot (s^{(51)}_{h_5 \sigma_1} \cdot s^{(52)}_{h_5 \sigma_2}) \cdot (s^{(63)}_{h_6 \sigma_3} \cdot s^{(64)}_{h_6 \sigma_4}). \tag{7}$$

The prime ideal I_f of polynomial invariants is generated by the 3×3-subdeterminants of the matrix

$$\begin{pmatrix} p_{0000} & p_{0010} & p_{0001} & p_{0011} \\ p_{0100} & p_{0110} & p_{0101} & p_{0111} \\ p_{1000} & p_{1010} & p_{1001} & p_{1011} \\ p_{1100} & p_{1110} & p_{1101} & p_{1111} \end{pmatrix} \tag{8}$$

Thus this particular model is the $k = l = 4$ instance of the determinantal variety in Proposition 4.

We generalize the determinantal presentation in this example by proposing the following explicit solution to Problem 5 for arbitrary binary trees τ. Every edge of τ induces a *split* of the set of leaves $\{1,2,\ldots,n\}$, corresponding to the two connected components of the tree obtained by removing that edge. The unrooted tree underlying τ is uniquely determined by the set of these splits.

Conjecture 13. *The ideal I_f of phylogenetic invariants of the general Markov model for any binary tree τ on binary random variables is generated by the 3×3-determinants of all two-dimensional matrices obtained by* flattening *the $2 \times \cdots \times 2$-table $(p_{\sigma_1 \cdots \sigma_n})$ according to the splits induced by the edges of τ.*

We need to explain the meaning of the word "flattening". If (A,B) is any split of the set $\{1,\ldots,n\}$ then this refers to the $2^{\#(A)} \times 2^{\#(B)}$-matrix whose rows and columns are indexed by functions $A \to \{0,1\}$ and $B \to \{0,1\}$ respectively, and whose entries are the 2^n probabilities $p_{\sigma_1 \cdots \sigma_n}$.

In December 2003, Allman and Rhodes announced a proof of the set-theoretic version of our Conjecture 13. What this means algebraically is that I_f equals the radical of the ideal generated by the aforementioned 3×3-determinants. In light of this progress, we wish to offer also the following tropical version of Conjecture 13. It would be very nice to show that Proposition 4 extends to this situation. However, none of the remaining discussion in this section depends on these conjectures.

Conjecture 14. *The map f is surjectively positive for $l = 2$. The tropical variety (resp. the positive tropical variety) of the prime ideal I_f coincides with the set of all $2 \times 2 \times \cdots \times 2$-tables $(u_{\sigma_1 \cdots \sigma_n})$ whose flattenings along the splits of the tree τ have tropical rank (resp. Barvinok rank) at most 2.*

The sum-product algorithm is used in practice to evaluate the polynomial (6). Its running time is linear in n, despite the fact that the number l^{n-1} of terms in (6) grows exponentially. This reduction in complexity is achieved by recursively grouping subsums. For instance, (7) becomes

$$p_{\sigma_1\sigma_2\sigma_3\sigma_4} \quad = \quad \sum_{\nu=0}^{1} \left(s^{(75)}_{\nu 0} s^{(51)}_{0\sigma_1} s^{(52)}_{0\sigma_2} + s^{(75)}_{\nu 1} s^{(51)}_{1\sigma_1} s^{(52)}_{1\sigma_2}\right) \cdot \left(s^{(76)}_{\nu 0} s^{(63)}_{0\sigma_3} s^{(64)}_{0\sigma_4} + s^{(76)}_{\nu 1} s^{(63)}_{1\sigma_3} s^{(64)}_{1\sigma_4}\right). \tag{9}$$

The rule to remember is this: Polynomials are evaluated recursively as sums of products of smaller polynomials. This is the solution to Problem 1. For details on the tree case see [10].

Problem 2 is known in phylogeny as the *joint ancestral reconstruction* problem, which asks for the maximum a posteriori ancestral assignments $\hat{h}_i$ given the observations $(\sigma_1,\ldots,\sigma_n)$ at the leaves. An efficient method for solving this problem appears in [16]. This method is nothing but the sum-product algorithm with ordinary arithmetic $(+,\times)$ replaced by tropical arithmetic $(\min,+)$. The σ-coordinate of the tropicalization $g : \mathbf{R}^{(2n-2)l^2} \to \mathbf{R}^{l^n}$ of the map (6) is

$$\delta_{\sigma_1 \cdots \sigma_n} \quad = \quad \min_h \sum_{\substack{i \in \tau' \\ \text{with children } j,k}} \left(v^{(ij)}_{h_i h_j} + v^{(ik)}_{h_i h_k}\right), \tag{10}$$

This expression can be evaluated efficiently by the same scheme as before. The rule now is this: Piecewise-linear concave functions are evaluated recursively as minima of sums of smaller such functions. A simple example illustrating this rule is the tropicalization of (9):

$$\delta_{\sigma_1\sigma_2\sigma_3\sigma_4} \quad = \quad \min_{\nu \in \{0,1\}} \left(u_{\nu\sigma_1\sigma_2} + u_{\nu\sigma_3\sigma_4}\right) \tag{11}$$

where $u_{\nu\sigma_1\sigma_2} = \min\left(v^{(75)}_{\nu 0} + v^{(51)}_{0\sigma_1} + v^{(52)}_{0\sigma_2}, v^{(75)}_{\nu 1} + v^{(51)}_{1\sigma_1} + v^{(52)}_{1\sigma_2}\right)$ and similarly for $u_{\nu\sigma_3\sigma_4}$.

We saw in Section 4 that the number of vertices of the Newton polytopes of the coordinate polynomials f_σ is critical for efficient parametric inference. That number grows polynomially in n if the number of parameters is fixed (thanks to Theorem 7) but it may grow exponentially if the number of parameters is not bounded. For the general Markov model on a tree τ, the growth will be exponential unless we restrict the number of parameters. This can be done, for instance, by considering the *homogeneous tree model* where the transition matrices along all edges are identical:

$$s^e_{\mu\nu} = s_{\mu\nu} \qquad \text{is independent of the edge } e.$$

Using Theorem 7, we obtain the following result analogous to Corollary 8.

Proposition 15. *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.*

For tree models which are used in applications, such as phylogenetics, the number of parameters is likely to be reduced even further. In such cases, the parametric joint ancestral reconstruction problem can be solved efficiently using the polytope propagation algorithm techniques in Proposition 11.

6 Summary: A Statistics – Geometry Dictionary

The algebraic representation for graphical models with hidden variables leads naturally to an interpretation of a parameterized model as a point on an algebraic variety. Marginal probabilities are coordinates of points on the variety. Varieties can be tropicalized, and the statistical meaning is that the MAP probabilities (calculated with logarithms of the parameters) can be interpreted as coordinates of points on the positive part of the tropical variety. Hence, the tropical model is fundamental for understanding MAP probabilities. Although we have not addressed it in this paper, the logarithms of the marginal probabilities are coordinates of points on the *amoeba* [23] of the model. Amoebas are likely to be important for understanding the geometry of maximum likelihood estimation.

The sum-product algorithm for graphical models is an efficient method for evaluating the coordinate polynomials of a graphical model. This algorithm works in exactly the same way for classical arithmetic $(+,\times)$ and for tropical arithmetic $(+,\min)$. This means that the same method is used to evaluate coordinates of points on the variety and of points on the tropical variety.

An explanation for an observation σ is a vertex of the Newton polytope of f_σ. Thus, the parametric inference problem is solved by finding the normal fans of the Newton polytopes of the coordinate polynomials. For many important applications, the number of vertices of the polytopes is polynomial in the size of the graphical model. The polytope propagation algorithm, which is a geometric analog of the sum-product algorithm, finds the Newton polytopes, and is efficient when the sum-product algorithm is fast and the number of vertices on the Newton polytopes is small.

An inference function for a graphical model is a function from the set of observations to the set of explanations which maximizes the a posteriori probabilities with respect to some choice of parameters. Inference functions correspond to vertices of the Newton polytope of the map f. This polytope is much larger than the Newton polytope of a single coordinate f_σ, so it can only be computed for small graphical models, but it has the advantage that it encodes the entire piecewise-linear geometry of the model.

In a companion paper [18], we show that polytope propagation is practical and useful in the important application of biological sequence analysis. In particular, existing parametric alignment methods [11, 13, 24] can be viewed as special cases of parametric inference for pair hidden Markov models. The computation of

the Newton polytopes is also useful for Bayesian computations, where we have priors on the parameters and it is of interest to integrate over the maximal cones in the normal fan of the Newton polytope [18, §5].

7 Acknowledgments

Lior Pachter was supported in part by a grant from the NIH (R01-HG02362-02). Bernd Sturmfels was supported by a Hewlett Packard Visiting Research Professorship 2003/2004 at MSRI Berkeley and in part by the NSF (DMS-0200729). We are grateful to Komei Fukuda, Michael Joswig and Kristian Ranestad for their help in obtaining the computational results reported in Section 2.

References

[1] G. Andrews: A lower bound for the volume of strictly convex bodies with many boundary points, Trans. Amer. Math. Soc. 106 (1963) 270–273.

[2] E. Allman and J. Rhodes: Phylogenetic invariants for the general Markov model of sequence mutation, Mathematical Biosciences, 186 (2003) 113–144.

[3] S. Aji and R. J. McEliece: The generalized distributive law, IEEE Transactions on Information Theory 46 (2000) 325–343.

[4] P. Baldi and S. Brunak: Bioinformatics. The Machine Learning Approach. A Bradford Book. The MIT Press. Cambridge, Massachusetts 1998.

[5] J. Cavender and J. Felsenstein: Invariants of phylogenies in a simple case with discrete states, Journal of Classification 4 (1987) 57–71.

[6] J. Cohen and U. Rothblum: Nonnegative ranks, decompositions, and factorizations of nonnegative matrices, Linear Algebra Appl. 190 (1993) 149–168.

[7] David Cox , Donal O'Shea and John Little: Ideals, Varieties and Algorithms, Springer Undergraduate Texts in Mathematics, 1996.

[8] M. Develin, F. Santos and B. Sturmfels: On the tropical rank of a matrix, preprint, `http://front.math.ucdavis.edu/math.CO/0312114`.

[9] M. Develin: The space of m points on a tree with n leaves, preprint. `http://front.math.ucdavis.edu/math.CO/0401224`.

[10] R. Durbin, S. Eddy, A. Krogh and G. Mitchison: Biological Sequence Analysis (Probabilistic Models of Proteins and Nucleic Acids), Cambridge University Press, 1998.

[11] D. Fernández-Baca, T. Seppäläinen and G. Slutzki: Parametric multiple sequence alignment and phylogeny construction, in Combinatorial Pattern Matching, Lecture Notes in Computer Science (R. Giancarlo and D. Sankoff eds.), Vol. 1848, 2000, 68–82.

[12] L.D. Garcia, M. Stillman and B. Sturmfels: Algebraic geometry of Bayesian networks, Journal of Symbolic Computation, to appear, `http://front.math.ucdavis.edu/math.AG/0301255`.

[13] D. Gusfield, K. Balasubramanian, and D. Naor: Parametric optimization of sequence alignment, Algorithmica 12 (1994) 312–326.

[14] F. Kschischang, B. Frey, and H. A. Loeliger: Factor graphs and the sum-product algorithm, IEEE Trans. Inform. Theory 47, Feb 2001, 498–519.

[15] M.I. Jordan and Y. Weiss: Graphical Models: Probabilistic Inference, in *Handbook of Brain Theory and Neural Networks, 2nd edition*, M. Arbib (Ed.), Cambridge, MA, MIT Press, 2002.

[16] T. Pupko, I. Pe'er, R. Shamir, and D. Graur: A fast algorithm for joint reconstruction of ancestral amino acid sequences, Molecular Biology and Evolution 17 (2000) 890–896.

[17] J. Richter-Gebert, B. Sturmfels, and T. Theobald: First steps in tropical geometry, in Idempotent Mathematics and Mathematical Physics (eds. G.L. Litvinov and V.P. Maslov), American Mathematical Society, 2004, preprint posted at `http://front.math.ucdavis.edu/math.AG/0306366`.

[18] L. Pachter and B. Sturmfels: Parametric inference for biological sequence analysis, companion paper, submitted.

[19] G. Pistone, E. Riccomagno, and H.P. Wynn: Algebraic Statistics: Computational Commutative Algebra in Statistics, Chapman and Hall, Boca Raton, Florida, 2001.

[20] L. R. Rabiner: A tutorial in hidden Markov models and selected applications in speech recognition, Proc. of the IEEE, 77 (1989) 257–286.

[21] D. Speyer and L. Williams: The tropical totally positive Grassmannian, preprint, `http://front.math.ucdavis.edu/math.CO/0312297`.

[22] B. Sturmfels: Gröbner Bases and Convex Polytopes, University Lecture Series, Vol. 8, American Mathematical Society, 1996.

[23] O. Viro: What is an amoeba? Notices of the American Math. Society 49 (2002) 916–917.

[24] M. Waterman, M. Eggert and E. Lander: Parametric sequence comparisons, Proc. Natl. Acad. Sci. USA 89 (1992) 6090–6093.

[25] D. West: Acyclic orientations of complete bipartite graphs, Discrete Math. 138 (1995) 393–396.