Formal Power Series and Algebraic Combinatorics
Séries Formelles et Combinatoire Algébrique
San Diego, California 2006

Bounds on the number of inference functions of a graphical model

Sergi Elizalde and Kevin Woods

Abstract. We give an upper bound on the number of inference functions of any directed graphical model. This bound is polynomial on the size of the model, for a fixed number of parameters, thus improving the exponential upper bound given in [Pachter and Sturmfels, Tropical Geometry of Statistical Models, *Proc. Natl. Acad. Sci.* 101, n. 46 (2004), 16132–16137]. Our proof reduces the problem to the enumeration of vertices of a Minkowski sum of polytopes. We also show that our bound is tight up to a constant factor, by constructing a family of hidden Markov models whose number of inference functions agrees asymptotically with the upper bound. Finally, we apply this bound to a model for sequence alignment that is used in computational biology.

Résumé. Nous donnons une limite supérieure sur le nombre de fonctions d'inférence de tout modéle graphique dirigé. Cette limite est polynômielle sur la grosseur du modèle, pour un nombre fixe de paramètres, améliorant ainsi la limite supérieure exponentielle donnée dans [Pachter and Sturmfels, Tropical Geometry of Statistical Models, *Proc. Natl. Acad. Sci.* 101, n. 46 (2004), 16132–16137]. Notre preuve réduit le problème à l'énumération de sommets d'une somme de Minkowski de polytopes. Nous montrons aussi que notre limite est serrée jusqu'à un facteur constant, en construisant une famille de modèles de Markov cachés dont le nombre de fonctions d'inférence coïncide asymptotiquement avec la limite supérieure. Finalement, nous appliquons cette limite à un modèle pour l'alignment de séquences qui est utilisé dans la biologie computationnelle.

1. Introduction

Many statistical models seek, given a set of observed data, to find the *hidden* (unobserved) data which best explains these observations. In this paper we consider graphical models, also called Bayesian networks, a broad class that includes many useful models, such as hidden Markov models (HMMs), pair hidden Markov models, and hidden tree models (background on graphical models will be given in Section 2.1). These graphical models relate the hidden and observed data probabilistically, and a natural problem is to determine, given a particular observation, what is the most likely hidden data (which is called the *explanation*). These models rely on parameters that are the probabilities relating the hidden and observed data. Any fixed values of the parameters determine a way to assign an explanation to each possible observation. This gives us a map, called an *inference function*, from observations to explanations.

An example of an inference function is the popular "*Did you mean*" feature from `google`, which could be implemented as a hidden Markov model, where the observed data is what we type into the computer, and the hidden data is what we were meaning to type. Graphical models are frequently used in these sorts of probabilistic approaches to artificial intelligence (see [**5**] for an introduction).

2000 *Mathematics Subject Classification.* Primary 62F15,52C45; Secondary 52B20,62P10,52B05.

Key words and phrases. inference functions, graphical models, sequence alignment, Newton polytope, normal fan.

Inference functions for graphical models are also important in computational biology [**6**, Section 1.5]. For example, consider the *gene-finding functions*, which were discussed in [**7**, Section 5]. These inference functions (corresponding to a particular HMM) are used to identify gene structures in DNA sequences. An observation in such a model is a sequence of nucleotides in the alphabet $\Sigma' = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, and an explanation is a sequence of 1's and 0's which indicate whether the particular nucleotide is in a gene or is not. We seek to use the information in the observed data (which we can find via DNA sequencing) to decide on the hidden information of which nucleotides are part of genes (which is hard to figure out directly). Another class of examples is that of sequence alignment models [**6**, Section 2.2]. In such models, an inference function is a map from a pair of DNA sequences to an optimal alignment of those sequences. If we change the parameters of the model, which alignments are optimal may change, and so the inference functions may change.

A surprising conclusion of this paper is that there cannot be *too many* different inference functions, though the parameters may vary continuously over all possible choices. For example, in the homogeneous binary HMM of length 5 (see Section 2.1 for some definitions; they are not important at the moment), the observed data is a binary sequence of length 5, and the explanation will also be a binary sequence of length 5. At first glance, there are

$$32^{32} = 1\ 461\ 501\ 637\ 330\ 902\ 918\ 203\ 684\ 832\ 716\ 283\ 019\ 655\ 932\ 542\ 976$$

possible maps from observed sequences to explanations. In fact, Christophe Weibel has computed that only 5266 of these possible maps are actually inference functions [**9**].

Different inference functions represent different criteria to decide what is the most likely explanation for each observation. A bound on the number of inference functions is important because it indicates how badly a model may respond to changes in the parameter values (which are generally known with very little certainty and only guessed at). Also, the polynomial bound given in Section 3 suggests that it might be feasible to precompute all the inference functions of a given graphical model, which would yield an efficient way to provide an explanation for each given observation.

This paper is structured as follows. In Section 2 we introduce some preliminaries about graphical models and inference functions, as well as some facts about polytopes. In Section 3 we present our main result. We call it the *Few Inference Functions Theorem*, and it states that in any graphical model the number of inference functions grows polynomially in the size of the model (if the number of parameters is fixed). The proof involves combinatorial tools, and a key step consists in reducing the enumeration of inference functions to the problem of counting the number of vertices of a certain polytope that is obtained as a Minkowski sum of smaller polytopes. In Section 4 we prove that our upper bound on the number of inference functions of a graphical model is sharp, up to a constant factor, by constructing a family of HMMs whose number of inference functions asymptotically matches the bound. In Section 5 we show that the bound is also asymptotically tight on a model for sequence alignment which is actually used in computational biology. In particular, this bound will be quadratic on the length of the input DNA sequences. We conclude with a few remarks and possible directions for further research.

2. Preliminaries

2.1. Graphical models. A *statistical model* is a family of joint probability distributions for a collection of discrete random variables $\mathbf{Z} = (Z_1, \ldots, Z_m)$, where each Z_i takes on values in some finite state space Σ_i. Here we will focus on directed graphical models. A *directed graphical model* (or *Bayesian network*) is a finite directed acyclic graph G where each vertex v_i corresponds to a random variable Z_i. Each vertex v_i also has an associated probability map

$$p_i : \left(\prod_{j:\ v_j \text{ a parent of } v_i} \Sigma_j \right) \to [0,1]^{|\Sigma_i|}.$$

Given the states of each Z_j such that v_j is a parent of v_i, the probability that v_i has a given state is independent of the values of all other vertices that are not descendants of v_i, and this map p_i gives that probability. In particular, we have the equality

$$\mathrm{Prob}(\mathbf{Z}=\tau)=\prod_i \mathrm{Prob}\left(Z_i=\tau_i,\ \text{given that } Z_j=\tau_j \text{ for all parents } v_j \text{ of } v_i\right)=\prod_i \left[p_i\left(\tau_{j_1},\ldots,\tau_{j_k}\right)\right]_{\tau_i},$$

where $v_{j_i},\ldots,v_{j_k}$ are the parents of v_i. Sources in the digraph (which have no parents) are generally given the uniform probability distribution on their states, though more general distributions are possible. See [**6**, Section 1.5] for general background on graphical models.

Example 2.1. The hidden Markov model (HMM) is a model with random variables $\mathbf{X}=(X_1,\ldots,X_n)$ and $\mathbf{Y}=(Y_1,\ldots,Y_n)$. Edges go from X_i to X_{i+1} and from X_i to Y_i.

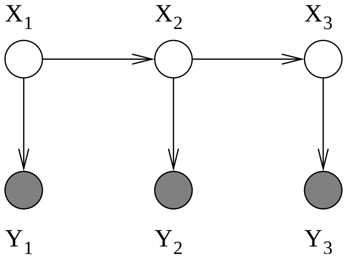

Figure 1. The graph of an HMM for $n=3$.

Generally, each X_i has the same state space Σ and each Y_i has the same state space Σ'. An HMM is called *homogeneous* if the p_{X_i}, for $1\le i\le n$ are identical and the p_{Y_i} are also identical. In this case, the p_{X_i} each correspond to the same $|\Sigma|\times|\Sigma|$ matrix $T=(t_{ij})$ (the *transition matrix*) and the p_{Y_i} each correspond to the same $|\Sigma|\times|\Sigma'|$ matrix $S=(s_{ij})$ (the *emission* matrix).

In the example, we have partitioned the variables into two sets. In general graphical models, we also have two kinds of variables: observed variables $\mathbf{Y}=(Y_1,Y_2,\ldots,Y_n)$ and hidden variables $\mathbf{X}=(X_1,X_2,\ldots,X_q)$. Generally, the observed variables are exactly the sinks of the directed graph, but this does not need to be the case. To simplify the notation, we make the assumption, which is often the case in practice, that all the observed variables take their values in the same finite alphabet Σ', and that all the hidden variables are on the finite alphabet Σ.

Notice that for given Σ and Σ' the homogeneous HMMs in this example depend only on a fixed set of parameters, t_{ij} and s_{ij}, even as n gets large. These are the sorts of models we are interested in. Let the number of parameters be a fixed integer d. We will name our parameters $\theta_1,\theta_2,\ldots,\theta_d$. By a graphical model with d parameters, we mean a graphical model such that each probability $\left[p_i\left(\tau_{j_1},\ldots,\tau_{j_k}\right)\right]_{\tau_i}$ is a monomial in our parameters, and furthermore the degree of this monomial is bounded by the number of parents of v_i. This is a natural assumption, because this probability is usually a product of one parameter for each edge incoming to v, as long as the parameters affect the probability of state v_i independently. This bound on degrees encompasses most interesting and useful graphical models. For example, in the homogeneous HMM, each v_i has only one parent, and the coordinates of p_i are degree one monomials (one of t_{ij} or s_{ij}).

In what follows we denote by E the number of edges of the underlying graph of a graphical model, by n the number of observed random variables, and by q the number of hidden random variables. The observations, then, are sequences in $(\Sigma')^n$ and the explanations are sequences in Σ^q. Let $l=|\Sigma|$ and $l'=|\Sigma'|$.

For each observation τ and hidden variables $\mathbf{h}$, $\mathrm{Prob}\left(\mathbf{X}=\mathbf{h},\ \mathbf{Y}=\tau\right)$ is a monomial $f_{\mathbf{h},\tau}$ of degree at most E in the parameters $\theta_1,\theta_2,\ldots,\theta_d$. Then for each observation $\tau\in(\Sigma')^n$, the observed probability $\mathrm{Prob}(\mathbf{Y}=\tau)$ is the sum over all hidden data $\mathbf{h}$ of $\mathrm{Prob}\left(\mathbf{X}=\mathbf{h},\ \mathbf{Y}=\tau\right)$, and so Prob($\mathbf{Y} = \tau$) is the polynomial $f_\tau = \sum_{\mathbf{h}} f_{\mathbf{h},\tau}$ in the parameters $\theta_1, \theta_2, \ldots, \theta_d$. The degree of f_τ is at most E.

Note that we have not assumed that the appropriate probabilities sum to 1. It turns out that the analysis is much easier if we do not place that restriction on our probabilities. At the end of the analysis, these restrictions may be added if desired (there are many models in use, however, which never place that restriction; these can no longer be properly called "probabilistic" models, but in fact belong to a more general class of "scoring" models which our analysis encompasses).

2.2. Inference functions. For fixed values of the parameters, the basic inference problem is to determine, for each given observation τ, the value $\mathbf{h} \in \Sigma^q$ of the hidden data that maximizes Prob($\mathbf{X} = \mathbf{h} \mid \mathbf{Y} = \tau$). A solution to this optimization problem is denoted $\widehat{\mathbf{h}}$ and is called an *explanation* of the observation τ. Each choice of parameter values $(\theta_1, \theta_2, \ldots, \theta_d)$ defines an *inference function* $\tau \mapsto \widehat{\mathbf{h}}$ from the set of observations $(\Sigma')^n$ to the set of explanations Σ^q.

It is possible that there is more than one value of $\widehat{\mathbf{h}}$ attaining the maximum of Prob($\mathbf{X} = \mathbf{h} \mid \mathbf{Y} = \tau$). In this case, for simplicity, we will pick only one such explanation, according to some consistent tie-breaking rule decided ahead of time. For example, we can pick the least such $\widehat{\mathbf{h}}$ in some given total order of the set Σ^q of hidden states. Another alternative would be to define inference functions as maps from $(\Sigma')^n$ to subsets of Σ^q. This would not affect the results of this paper, so for the sake of simplicity, we consider only inference functions as defined above.

It is interesting to observe that the total number of maps $(\Sigma')^n \longrightarrow \Sigma^q$ is $(l^q)^{(l')^n} = l^{q(l')^n}$, which is doubly-exponential in the length n of the observations. However, most of these maps are not inference functions for any values of the parameters. Before our results, the best upper bound in the literature was an exponential bound given in [**8**, Corollary 10]. In Section 3 we give a polynomial upper bound on the number of inference functions of a graphical model.

2.3. Polytopes. Here we review some facts about convex polytopes, and we introduce some notation. Recall that a polytope is a bounded intersection of finitely many closed halfspaces, or equivalently, the convex hull of a finite set of points. For the basic definitions about polytopes we refer the reader to [**10**].

Given a polynomial $f(\theta) = \sum_{i=1}^N \theta_1^{a_{1,i}} \theta_2^{a_{2,i}} \cdots \theta_d^{a_{d,i}}$, its *Newton polytope*, denoted by NP(f), is defined as the convex hull in $\mathbb{R}^d$ of the set of points $\{(a_{1,i}, a_{2,i}, \ldots, a_{d,i}) : i = 1, \ldots, N\}$. For example, if $f(\theta_1, \theta_2) = 2\theta_1^3 + 3\theta_1^2\theta_2^2 + \theta_1\theta_2^2 + 3\theta_1 + 5\theta_2^4$, then its Newton polytope NP(f) is given in Figure 2.

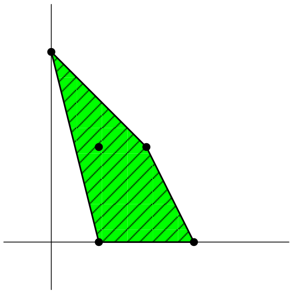

Figure 2. The Newton polytope of $f(\theta_1, \theta_2) = 2\theta_1^3 + 3\theta_1^2\theta_2^2 + \theta_1\theta_2^2 + 3\theta_1 + 5\theta_2^4$.

Given a polytope $P \subset \mathbb{R}^d$ and a vector $w \in \mathbb{R}^d$, the set of all points in P at which the linear functional $x \mapsto x \cdot w$ attains its maximum determines a *face* of P. It is denoted

$$\mathrm{face}_w(P) \quad = \quad \big\{ x \in P \ : \ x \cdot w \geq y \cdot w \ \text{ for all } y \in P \big\}.$$

Faces of dimension 0 (consisting of a single point) are called *vertices*, and faces of dimension 1 are called *edges*. If d is the dimension of the polytope, then faces of dimension $d-1$ are called *facets*.

Let P be a polytope and F a face of P. The *normal cone* of P at F is

$$N_P(F) \quad = \quad \left\{ w \in \mathbb{R}^d \,:\, \mathrm{face}_w(P) = F \right\}.$$

The collection of all cones $N_P(F)$ as F runs over all faces of P is denoted $\mathcal{N}(P)$ and is called the *normal fan* of P. Thus the normal fan $\mathcal{N}(P)$ is a partition of $\mathbb{R}^d$ into cones. The cones in $\mathcal{N}(P)$ are in bijection with the faces of P, and if $w \in N_P(F)$ then the linear functional $x \cdot w$ is maximized on F. Figure 3 shows the normal fan of a polytope.

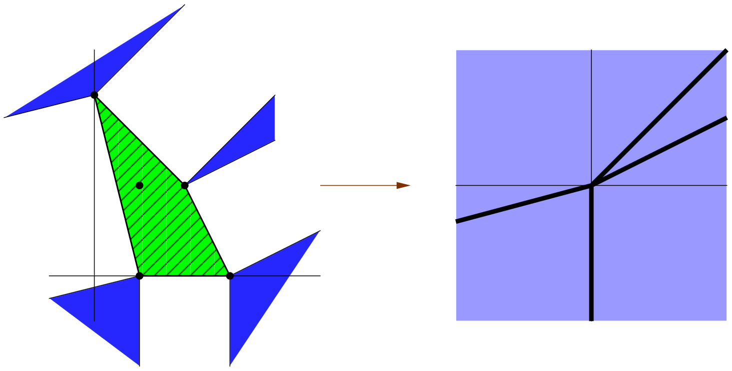

FIGURE 3. The normal fan of a polytope.

The *Minkowski sum* of two polytopes P and P' is defined as

$$P + P' := \{\mathbf{x} + \mathbf{x}' : \mathbf{x} \in P,\, \mathbf{x}' \in P'\}.$$

The *common refinement* of two or more normal fans is the collection of cones obtained as the intersection of a cone from each of the individual fans. For polytopes $P_1, P_2, \ldots, P_k$, the common refinement of their normal fans is denoted $\mathcal{N}(P_1) \wedge \cdots \wedge \mathcal{N}(P_k)$. The following lemma states the well-known fact that the normal fan of a Minkowski sum of polytopes is the common refinement of their individual fans (see [**10**, Proposition 7.12] or [**2**, Lemma 2.1.5]):

LEMMA 2.2. *$\mathcal{N}(P_1 + \cdots + P_k) = \mathcal{N}(P_1) \wedge \cdots \wedge \mathcal{N}(P_k)$.*

We finish with a result of Gritzmann and Sturmfels that will be useful later. It gives a bound on the number of vertices of a Minkowski sum of polytopes.

THEOREM 2.3 ([**2**]). *Let $P_1, P_2, \ldots, P_k$ be polytopes in $\mathbb{R}^d$, and let m denote the number of non-parallel edges of $P_1, \ldots, P_k$. Then the number of vertices of $P_1 + \cdots + P_k$ is at most*

$$2 \sum_{j=0}^{d-1} \binom{m-1}{j}.$$

Note that this bound is independent of the number k of polytopes.

3. An upper bound on the number of inference functions

For fixed parameters, the inference problem of finding the explanation $\widehat{\mathbf{h}}$ that maximizes $\mathrm{Prob}(\mathbf{X} = \mathbf{h} \mid \mathbf{Y} = \tau)$ is equivalent to identifying the monomial $f_{\mathbf{h},\tau} = \theta_1^{a_{1,\mathbf{h}}} \theta_2^{a_{2,\mathbf{h}}} \cdots \theta_d^{a_{d,\mathbf{h}}}$ of f_τ with maximum value. Since the logarithm is a monotonically increasing function, the desired monomial also maximizes the quantity

$$\begin{aligned} \log(\theta_1^{a_{1,\mathbf{h}}} \theta_2^{a_{2,\mathbf{h}}} \cdots \theta_d^{a_{d,\mathbf{h}}}) &= a_{1,\mathbf{h}} \log(\theta_1) + a_{2,\mathbf{h}} \log(\theta_2) + \cdots + a_{d,\mathbf{h}} \log(\theta_d) \\ &= a_{1,\mathbf{h}} v_1 + a_{2,\mathbf{h}} v_2 + \cdots + a_{d,\mathbf{h}} v_d, \end{aligned}$$

where we replace $\log(\theta_i)$ with v_i. This is equivalent to the fact that the corresponding point $(a_{1,\mathbf{h}}, a_{2,\mathbf{h}}, \ldots, a_{d,\mathbf{h}})$ maximizes the linear expression $v_1x_1 + \cdots + v_dx_d$ on the Newton polytope $\mathrm{NP}(f_\tau)$. Thus, the inference problem for fixed parameters becomes a linear programming problem.

Each choice of the parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ determines an inference function. If $\mathbf{v} = (v_1, v_2, \ldots, v_d)$ is the vector in $\mathbb{R}^d$ with coordinates $v_i = \log(\theta_i)$, then we denote the corresponding inference function by

$$\Phi_{\mathbf{v}} : (\Sigma')^n \longrightarrow \Sigma^q.$$

For each observation $\tau \in (\Sigma')^n$, its explanation $\Phi_{\mathbf{v}}(\tau)$ is given by the vertex of $\mathrm{NP}(f_\tau)$ that is maximal in the direction of the vector $\mathbf{v}$. Note that for certain values of the parameters (if $\mathbf{v}$ is perpendicular to a positive-dimensional face of $\mathrm{NP}(f_\tau)$) there may be more than one vertex attaining the maximum. It is also possible that a single point $(a_{1,\mathbf{h}}, a_{2,\mathbf{h}}, \ldots, a_{d,\mathbf{h}})$ in the polytope corresponds to several different values of the hidden data. In both cases, we pick the explanation according to the tie-breaking rule determined ahead of time. This simplification does not affect the asymptotic number of inference functions.

Different values of θ yield different directions $\mathbf{v}$, which can result in distinct inference functions. We are interested in bounding the number of different inference functions that a graphical model can have. The next theorem gives an upper bound which is polynomial in the size of the graphical model. In fact, very few of the $l^{q(l')^n}$ functions $(\Sigma')^n \longrightarrow \Sigma^q$ are inference functions.

THEOREM 3.1 (The Few Inference Functions Theorem). *Let d be a fixed positive integer. Consider a graphical model with d parameters, and let E be the number of edges of the underlying graph. Then, the number of inference functions of the model is at most $O(E^{d(d-1)})$.*

Before proving this theorem, observe that the number E of edges depends on the number n of observed random variables. In most graphical models of interest, E is a linear function of n, so the bound becomes $O(n^{d(d-1)})$. For example, the hidden Markov model has $E = 2n - 1$ edges. The only property of the number E that we actually need in the proof is that it is a bound on the degrees of the monomials $f_{\mathbf{h},\tau}$.

PROOF. In the first part of the proof we will reduce the problem of counting inference functions to the enumeration of the vertices of a certain polytope. We have seen that an inference function is specified by a choice of the parameters, which is equivalent to choosing a vector $\mathbf{v} \in \mathbb{R}^d$. The function is denoted $\Phi_{\mathbf{v}} : (\Sigma')^n \longrightarrow \Sigma^q$, and the explanation $\Phi_{\mathbf{v}}(\tau)$ of a given observation τ is determined by the vertex of $\mathrm{NP}(f_\tau)$ that is maximal in the direction of $\mathbf{v}$. Thus, cones of the normal fan $\mathcal{N}(\mathrm{NP}(f_\tau))$ correspond to sets of vectors $\mathbf{v}$ that give rise to the same explanation for the observation τ. Non-maximal cones (i.e., those contained in another cone of higher dimension) correspond to directions $\mathbf{v}$ for which more than one vertex is maximal. Since ties are broken using a consistent rule, we disregard this case for simplicity. Thus, in what follows we consider only maximal cones of the normal fan.

Let $\mathbf{v}' = (v'_1, v'_2, \ldots, v'_d)$ be another vector corresponding to a different choice of parameters (see Figure 4). By the above reasoning, $\Phi_{\mathbf{v}}(\tau) = \Phi_{\mathbf{v}'}(\tau)$ if and only if $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{N}(\mathrm{NP}(f_\tau))$. Thus, $\Phi_{\mathbf{v}}$ and $\Phi_{\mathbf{v}'}$ are the same inference function if and only if $\mathbf{v}$ and $\mathbf{v}'$ belong to the same cone of $\mathcal{N}(\mathrm{NP}(f_\tau))$ for all observations $\tau \in (\Sigma')^n$. Consider the common refinement of all these normal fans, $\bigwedge_{\tau\in(\Sigma')^n} \mathcal{N}(\mathrm{NP}(f_\tau))$. Then, $\Phi_{\mathbf{v}}$ and $\Phi_{\mathbf{v}'}$ are the same inference function exactly when $\mathbf{v}$ and $\mathbf{v}'$ lie in the same cone of this common refinement. This implies that the number of inference functions equals the number of cones in $\bigwedge_{\tau\in(\Sigma')^n} \mathcal{N}(\mathrm{NP}(f_\tau))$. By Lemma 2.2, this common refinement is the normal fan of $\mathrm{NP}(\mathbf{f}) = \sum_{\tau\in(\Sigma')^n} \mathrm{NP}(f_\tau)$, the Minkowski sum of the polytopes $\mathrm{NP}(f_\tau)$ for all observations τ. It follows that enumerating inference functions is equivalent to counting vertices of $\mathrm{NP}(\mathbf{f})$. In the remaining part of the proof we give an upper bound on the number of vertices of $\mathrm{NP}(\mathbf{f})$.

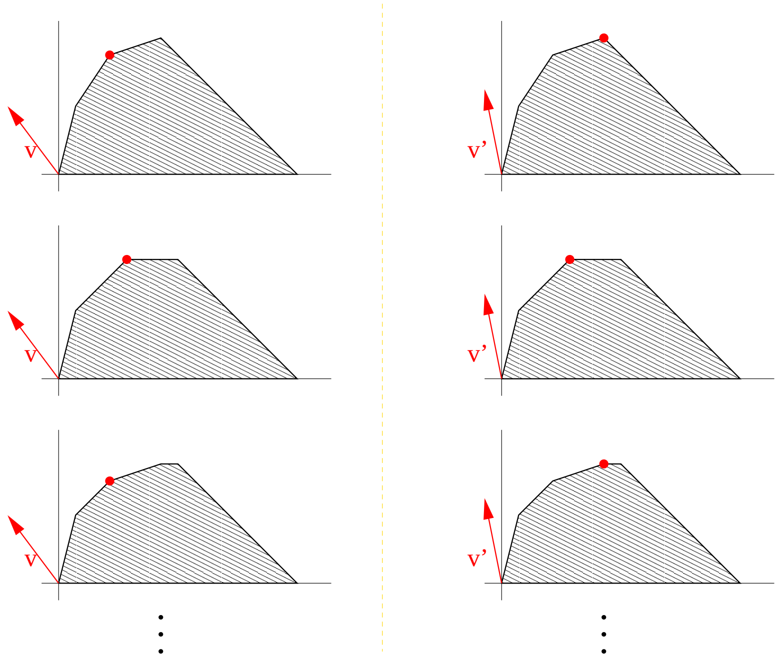

FIGURE 4. Two different inference functions, $\Phi_{\mathbf{v}}$ (left column) and $\Phi_{\mathbf{v}'}$ (right column). In each row is the Newton polytope corresponding to a different observation. The respective explanations are given by the marked vertices in each polytope.

Note that for each τ, the polytope $\mathrm{NP}(f_\tau)$ is contained in the hypercube $[0,E]^d$, since each parameter θ_i can appear as a factor of a monomial of f_τ at most E times. Also, the vertices of $\mathrm{NP}(f_\tau)$ have integral coordinates, because they are exponent vectors. Polytopes whose vertices have integral coordinates are called *lattice polytopes*. It follows that the edges of $\mathrm{NP}(f_\tau)$ are given by vectors where each coordinate is an integer between $-E$ and E. There are only $(2E+1)^d$ such vectors, so this is an upper bound on the number of different directions that the edges of the polytopes $\mathrm{NP}(f_\tau)$ can have.

This property of the Newton polytopes of the coordinates of the model will allow us to give an upper bound on the number of vertices of their Minkowski sum $\mathrm{NP}(\mathbf{f})$. The last ingredient that we need is Theorem 2.3. In our case we have a sum of polytopes $\mathrm{NP}(f_\tau)$, one for each observation $\tau \in (\Sigma')^n$, having at most $(2E+1)^d$ non-parallel edges in total. Hence, by Theorem 2.3, the number of vertices of $\mathrm{NP}(\mathbf{f})$ is at most

$$2\sum_{j=0}^{d-1}\binom{(2E+1)^d-1}{j}.$$

As E goes to infinity, the dominant term of this expression is

$$\frac{2^{d^2-d+1}}{(d-1)!}\,E^{d(d-1)}.$$

Thus, we get an $O(E^{d(d-1)})$ upper bound on the number of inference functions of the graphical model. □

In the next section we show that the bound given in Theorem 3.1 is tight up to a constant factor.

4. A lower bound

As before, we fix d, the number of parameters in our model. The Few Inference Functions Theorem (Theorem 3.1) tells us that the number of inference functions is bounded from above by some function $cE^{d(d-1)}$, where c is a constant (depending only on d) and E is the number of edges in the graphical model. Here we show that this bound is tight up to a constant, by constructing a family of graphical models whose number of inference functions is at least $\tilde{c}E^{d(d-1)}$, where $\tilde{c}$ is another constant. In fact, we will construct a family of hidden Markov models with this property. To be precise, we have the following theorem.

THEOREM 4.1. *Fix d. There is a constant $c' = c'(d)$ such that, given $n \in \mathbb{Z}_+$, there exists an HMM of length n, with d parameters, $2d+2$ hidden states, and 2 observed states, such that there are at least $c'n^{d(d-1)}$ distinct inference functions. (For HMMs, $E = 2n-1$, so this also gives us the lower bound in terms of E).*

In the proof of this theorem, we will state several lemmas that must be used. We omit the proofs of some of them here due to lack of space. Given n, we first construct the appropriate HMM, $\mathcal{M}_n$, using the following lemma.

LEMMA 4.2. *Given $n \in \mathbb{Z}_+$, there is an HMM, $\mathcal{M}_n$, of length n, with d parameters, $2d+2$ hidden states, and 2 observed states, such that for any $a = (a_1, \ldots, a_n) \in \mathbb{Z}_+^d$ with $\sum_i a_i < n$, there is an observed sequence which has one explanation if*

$$a_1 \log(\theta_1) + \cdots + a_d \log(\theta_d) > 0$$

and another explanation if

$$a_1 \log(\theta_1) + \cdots + a_d \log(\theta_d) < 0.$$

PROOF. Given d and n, define a length n HMM with parameters $\theta_1, ..., \theta_d$, as follows. The observed states will be S and C (for "start of block," and "continuing block," respectively). The hidden states will be s_i, s'_i, c_i, and c'_i, for $1 \leq i \leq d+1$ (think of s_i and s'_i as "start of the ith block" and c_i and c'_i as "continuing the ith block").

Here's the idea of what we want this HMM to do: if the observed sequence has S's in position 1, a_1+1, a_1+a_2+1, …, and $a_1+\cdots+a_d+1$ and C's elsewhere, then there will be only two possibilities for the sequence of hidden states, either

$$t = s_1 \underbrace{c_1 \cdots c_1}_{a_1-1} s_2 \underbrace{c_2 \cdots c_2}_{a_2-1} \cdots s_d \underbrace{c_d \cdots c_d}_{a_d-1} s_{d+1} \underbrace{c_{d+1} \cdots c_{d+1}}_{n-a_1-\cdots-a_d-1}$$

or

$$t' = s'_1 \underbrace{c'_1 \cdots c'_1}_{a_1-1} s'_2 \underbrace{c'_2 \cdots c'_2}_{a_2-1} \cdots s'_d \underbrace{c'_d \cdots c'_d}_{a_d-1} s'_{d+1} \underbrace{c'_{d+1} \cdots c'_{d+1}}_{n-a_1-\cdots-a_d-1}.$$

We will also make sure that t has a priori probability

$$\theta_1^{a_1} \cdots \theta_d^{a_d}$$

and t' has a priori probability 1. Then t is the explanation if $a_1 \log(\theta_1) + \cdots + a_d \log(\theta_d) > 0$ and t' is the explanation if $a_1 \log(\theta_1) + \cdots + a_d \log(\theta_d) < 0$. Remember that we are not constraining our probability sums to be 1. A very similar HMM could be constructed that obeys that constraint, if desired. But to simplify notation it will be more convenient to treat the transition probabilities as parameters that do not necessarily sum to one at each vertex, even if this forces us to use the term "probability" somewhat loosely.

Here is how we set up the transitions/emmisions. Let s_i and s'_i, for $1 \leq i \leq d+1$, all emit S with probability 1 and C with probability 0. Let c_i and c'_i emit C with probability 1 and S with probability 0. Let s_i, for $1 \leq i \leq d$, transition to c_i with probability θ_i and transition to everything else with probability 0. Let s_{d+1} transition to c_{d+1} with probability 1 and to everything else with probability 0. Let s_i', for $1 \leq i \leq d+1$, transition to c_i' with probability 1 and to everything else with probability 0. Let c_i, for $1 \leq i \leq d$, transition to c_i with probability θ_i, to s_{i+1} with probability θ_i, and to everything else with probability 0. Let c_{d+1} transition to c_{d+1} with probability 1, and to everything else with probability 0. Let c_i', for $1 \leq i \leq d$ transition to c_i' with probability 1, to s_{i+1} with probability 1, and to everything else with probability 0. Let c_{d+1}' transition to c_{d+1}' with probability 1 and to everything else with probability 0.

Starting with the uniform probability distribution on the first hidden state, this does exactly what we want it to: given the correct observed sequence, t and t' are the only explanations, with the correct probabilities. □

This means that, for the HMM provided by this lemma, the decomposition of (log-)parameter space into inference cones includes all of the hyperplanes $\{x : \langle a, x\rangle = 0\}$ such that $a \in \mathbb{Z}_+^d$ with $\sum_i a_i < n$. Call the arrangement of these hyperplanes $\mathcal{H}_n$. It suffices to show that the arrangement $\mathcal{H}_n$ consists of at least $c'n^{d(d-1)}$ chambers (full dimensional cones determined by the arrangement). There are $c_1 n^d$ ways to choose one of the hyperplanes from $\mathcal{H}_n$, for some constant c_1. Therefore there are $c_1^{d-1} n^{d(d-1)}$ ways to choose $d-1$ of the hyperplanes; their intersection is, in general, a 1-dimensional face of $\mathcal{H}_n$ (that is, the intersection is a ray which is an extreme ray for the cones it is contained in). It is quite possible that two different ways of choosing $d-1$ hyperplanes give the same extreme ray. The following lemma says that some constant fraction of these choices of extreme rays are actually distinct.

Lemma 4.3. *Fix d. Given n, let $\mathcal{H}_n$ be the hyperplane arrangement consisting of the hyperplanes of the form $\{x : \langle a, x\rangle = 0\}$ with $a \in \mathbb{Z}_+^d$ and $\sum_i a_i < n$. Then the number of 1-dimensional faces of $\mathcal{H}_n$ is $c_2 n^{d(d-1)}$, for some constant c_2.*

Each chamber will have a number of these extreme rays on its boundary. The following lemma gives a constant bound on this number.

Lemma 4.4. *Fix d. Given n, define $\mathcal{H}_n$ as above. Each chamber of $\mathcal{H}_n$ has at most $2^{d(d-1)}$ extreme rays.*

Conversely, each ray is an extreme ray for at least 1 chamber. Therefore there are at least $\frac{c_2}{2^{d(d-1)}} n^{d(d-1)}$ chambers, and Theorem 4.1 is proved.

5. Inference functions for sequence alignment

In this section we give an application of Theorem 3.1 to a basic model for sequence alignment. Sequence alignment is one of the most frequently used techniques in determining the similarity between biological sequences. In the standard instance of the sequence alignment problem, we are given two sequences (usually DNA or protein sequences) that have evolved from a common ancestor via a series of mutations, insertions and deletions. The goal is to find the best alignment between the two sequences. The definition of "best" here depends on the choice of scoring scheme, and there is often disagreement about the correct choice. In *parametric sequence alignment*, this problem is circumvented by instead computing the optimal alignment as a function of *variable* scores. Here we consider one such scheme, in which all matches are equally rewarded, all mismatches are equally penalized and all spaces are equally penalized. Efficient parametric sequence alignment algorithms are known (see for example [**6**, Chapter 7]). Here we are concerned with the different inference functions that car arise when the parameters vary. For a detailed treatment on the subject of sequence alignment, we refer the reader to [**3**].

Given two strings σ^1 and σ^2 of lengths n_1 and n_2 respectively, an *alignment* is a pair of equal length strings (μ^1, μ^2) obtained from σ^1, σ^2 by inserting dashes "$-$" in such a way that there is no position in which both μ^1 and μ^2 have a dash. A *match* is a position where μ^1 and μ^2 have the same character, a *mismatch* is a position where μ^1 and μ^2 have different characters, and a *space* is a position in which one of μ^1 and μ^2 has a dash. A simple scoring scheme consists of two parameters α and β denoting mismatch and space penalties respectively. The reward of a match is set to 1. The score of an alignment with z matches, x mismatches, and y spaces is then $z - x\alpha - y\beta$. Observe that these numbers always satisfy $2z + 2x + y = n_1 + n_2$.

This model for sequence alignment is a particular case of a so-called pair hidden Markov model. The problem of determining the highest scoring alignment for given values of α and β is equivalent to the inference problem in the pair hidden Markov model. In this setting, an observation is a pair of sequences $\tau = (\sigma^1, \sigma^2)$, and the number of observed variables is $n = n_1 + n_2$. The values of the hidden variables in an explanation indicate the positions of the spaces in the optimal alignment. We will refer to this as the *2-parameter model for sequence alignment.*

For each pair of sequences τ, the Newton polytope of the polynomial f_τ is the convex hull of the points (x, y, z) whose coordinates are the number of mismatches, spaces, and matches, respectively, of each possible alignment of the pair. This polytope is only two dimensional, as it lies on the plane $2z+2x+y = n_1+n_2$. No information is lost by considering its projection onto the xy-plane instead. This projection is just the convex hull of the points (x, y) giving the number of mismatches and spaces of each alignment. For any alignment of sequences of lengths n_1 and n_2, the corresponding point (x, y) lies inside the square $[0, n]^2$, where $n = n_1 + n_2$. Therefore, since we are dealing with lattice polygons inside $[0, n]^2$, it follows from the proof of the Few Inference Functions Theorem (Theorem 3.1) that the number of inference functions of this model is $O(n^{2(2-1)})) = O(n^2)$. Next we show that this quadratic bound is tight, even in the case of the binary alphabet.

PROPOSITION 5.1. *Consider the 2-parameter model for sequence alignment for two observed sequences of length n and let $\Sigma' = \{0, 1\}$ be the binary alphabet. Then, the number of inference functions of this model is $\Theta(n^2)$.*

PROOF. The above argument shows that $O(n^2)$ is an upper bound on the number of inference functions of the model. To prove the proposition, we will argue that there are at least $\Omega(n^2)$ such functions.

Since the two sequences have the same length, the number of spaces in any alignment is even. For convenience, we define $y' = y/2$ and $\beta' = 2\beta$, and we will work with the coordinates (x, y', z) and the parameters α and β'. The value y' is called the number of insertions (half the number of spaces), and β' is the insertion penalty. For fixed values of α and β', the explanation of an observation $\tau = (\sigma^1, \sigma^2)$ is given by the vertex of $\mathrm{NP}(f_\tau)$ that is maximal in the direction of the vector $(-\alpha, -\beta', 1)$. In this model, $\mathrm{NP}(f_\tau)$ is the convex hull of the points (x, y', z) whose coordinates are the number of mismatches, insertions and matches of the alignments of σ^1 and σ^2.

The argument in the proof of Theorem 3.1 shows that the number of inference functions of this model is the number of cones in the common refinement of the normal fans of $\mathrm{NP}(f_\tau)$, where τ runs over all pairs of sequences of length n in the alphabet Σ'. Since the polytopes $\mathrm{NP}(f_\tau)$ lie on the plane $x + y' + z = n$, it is equivalent to consider the normal fans of their projections onto the $y'z$-plane. These projections are lattice polygons contained in the square $[0, n]^2$. We denote by P_τ the projection of $\mathrm{NP}(f_\tau)$ onto the $y'z$-plane.

We will construct, for any relatively prime positive integers u and v with $u < v$ and $6v - 2u \le n$, a pair $\tau = (\sigma^1, \sigma^2)$ of binary sequences of length n such that P_τ has an edge of slope u/v. Such an edge gives rise to the line $u \cdot \alpha + v \cdot \beta' = 0$ separating regions in the normal fan $\mathcal{N}(P_\tau)$ and hence in $\bigwedge_\tau \mathcal{N}(P_\tau)$, where τ ranges over all pairs of binary sequences of length n. The number of such choices u, v is $\Omega(n^2)$ (this relies on the fact, see [**1**, Chapter 3], that a positive fraction of choices of $(u, v) \in \mathbb{Z}^2$ have u and v relatively prime). This implies that the number of different inference functions is $\Omega(n^2)$.

Thus, it only remains to construct such a τ, given positive integers u and v as above. Let $a := 2v$, $b := v - u$. Assume first that $n = 6v - 2u = 2a + 2b$. Consider the sequences $\sigma^1 = 0^a 1^b 0^b 1^a$, $\sigma^2 = 1^a 0^b 1^b 0^a$, where 0^a indicates that the symbol 0 is repeated a times. Let $\tau = (\sigma^1, \sigma^2)$. Then, it is not hard to see that the polygon P_τ for this pair of sequences has four vertices: $v_0 = (0, 0)$, $v_1 = (b, 3b)$, $v_2 = (a+b, a+b)$ and $v_3 = (n, 0)$. The slope of the edge between v_1 and v_2 is $\frac{a-2b}{a} = \frac{u}{v}$.

If $n > 6v - 2u = 2a + 2b$, we just append $0^{n-2a-2b}$ to both sequences σ^1 and σ^2. In this case, the vertices of P_τ are $(0, n - 2a - 2b)$, $(b, n - 2a + b)$, $(a + b, n - a - b)$, $(n, 0)$ and $(n - 2a - 2b, 0)$.

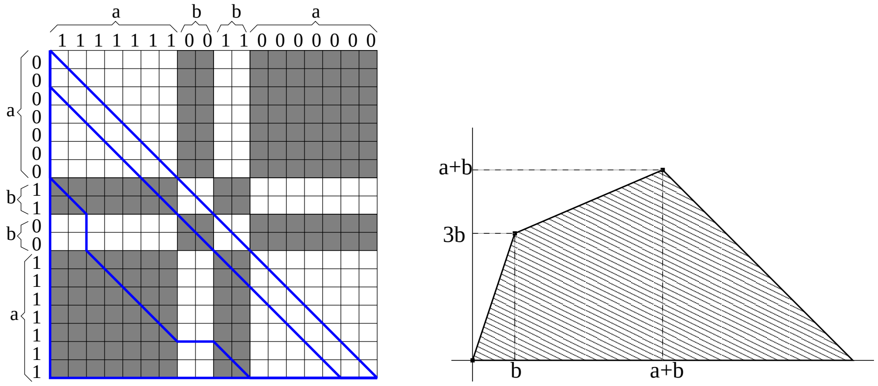

FIGURE 5. A pair of binary sequences of length 18 giving the slope 3/7 in their alignment polytope. The four paths in the *alignment graph* on the left correspond to the four vertices; a right step in the graph corresponds to a space in σ^1, a down step to a space in σ^2, and a diagonal step to a match or mismatch. See [**6**, Section 2.2] for a full definition of the alignment graph.

Note that if $v - u$ is even, the construction can be done with sequences of length $n = 3v - u$ by taking $a := v$, $b := \frac{v-u}{2}$. Figure 5 shows the alignment graph and the polygon P_τ for $a = 7$, $b = 2$. □

In most cases, one is interested only in those inference functions that are biologically meaningful. This corresponds to parameter values with $\alpha, \beta \geq 0$, which means that mismatches and spaces are penalized instead of rewarded. Sometimes one also requires that $\alpha \leq \beta$, which means that a mismatch should be penalized less than two spaces. It is interesting to observe that our construction in the proof of Proposition 5.1 not only shows that the total number of inference functions is $\Omega(n^2)$, but also that the number of biologically meaningful ones is still $\Omega(n^2)$. This is because the different rays created in our construction have a biologically meaningful direction in the parameter space.

6. Final remarks

An interpretation of Theorem 3.1 is that the ability to change the values of the parameters of a graphical model does not give as much freedom as it may appear. There is a very large number of possible ways to assign an explanation to each observation. However, only a tiny proportion of these come from a consistent method for choosing the most probable explanation for a certain choice of parameters. Even though the parameters can vary continuously, the number of different inference functions that can be obtained is at most polynomial in the number of edges of the model, assuming that the number of parameters is fixed.

In the case of sequence alignment, the number of possible functions that associate an alignment to each pair of sequences of length n is doubly-exponential in n. However, the number of functions that pick the alignment with highest score in the 2-parameter model, for some choice of the parameters α and β, is only $\Theta(n^2)$. Thus, most ways of assigning alignments to pairs of sequences do not correspond to any consistent choice of parameters. If we use a model with more parameters, say d, the number of inference functions may be larger, but still polynomial in n, namely $O(n^{d(d-1)})$.

Having shown that the number of inference functions of a graphical model is polynomial in the size of the model, an interesting next step would be to find an efficient way to precompute all the inference functions for given models. This would allow us to give the answer (the explanation) to a query (an observation) very quickly. Theorem 3.1 suggests that it might be computationally feasible to precompute the polytope NP($\mathbf{f}$), whose vertices correspond to the inference functions. However, the difficulty arises when we try to describe a particular inference function efficiently. The problem is that the characterization of an inference function involves an exponential number of observations.

References

[1] T.M. Apostol, *Introduction to Analytic Number Theory*, Springer-Verlag, New York, 1976.

[2] P. Gritzmann, B. Sturmfels, Minkowski addition of polytopes: Computational complexity and applications to Gröbner bases, *SIAM Journal of Discrete Mathematics* 6 (1993), 246–269.

[3] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press, 1997.

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

[5] F. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.

[6] L. Pachter, B. Sturmfels, editors, *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.

[7] L. Pachter, B. Sturmfels, The Mathematics of Phylogenomics, submitted.

[8] L. Pachter, B. Sturmfels, Tropical Geometry of Statistical Models, *Proc. Natl. Acad. Sci.* 101, n. 46 (2004), 16132–16137.

[9] Christophe Weibel, personal commnuication.

[10] G.M. Ziegler, *Lectures on Polytopes*, Graduate Texts in Mathematics 152, Springer, New York, 1995.

DEPARTMENT OF MATHEMATICS, DARTMOUTH COLLEGE, HANOVER, NH 03755
E-mail address: sergi.elizalde@dartmouth.edu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, BERKELEY, CA 94720
E-mail address: kwoods@math.berkeley.edu