

Binary Hidden Markov Models and Varieties

Andrew Critch, UC Berkeley*

June 5, 2012

Abstract

The technological applications of hidden Markov models have been extremely diverse and successful, including natural language processing, gesture recognition, gene sequencing, and Kalman filtering of physical measurements. HMMs are highly non-linear statistical models, and just as linear models are amenable to linear algebraic techniques, non-linear models are amenable to commutative algebra and algebraic geometry.

This paper examines closely those HMMs in which all the random variables, called nodes, are binary. Its main contributions are (1) *minimal defining equations* for the 4-node model, comprising 21 quadrics and 29 cubics, which were computed using Gröbner bases in the cumulant coordinates of Sturmfels and Zwiernik, and (2) a *birational parametrization* for every binary HMM, with an explicit inverse for recovering the hidden parameters in terms of observables. The new model parameters in (2) are hence rationally identifiable in the sense of Sullivant, Garcia-Puente, and Spielvogel, and each model's Zariski closure is therefore a *rational* projective variety of dimension 5. Gröbner basis computations for the model and its graph are found to be considerably faster using these parameters. Together, (1) and (2) provide a nearly instantaneous computational test for whether an observed probability distribution is due to a binary hidden Markov process, in comparison with a less specialized algorithm of Schönhuth involving matrix row reduction. Defining equations such as (1) have been used successfully in model selection problems in phylogenetics, and one can hope for similar applications in the case of HMMs.

1 Introduction

The present work is motivated primarily by the problems of *model selection* and *parameter identifiability*, viewed from the perspective of algebraic geometry. By beginning with the simplest hidden Markov models (HMMs) — those with all nodes binary — the hope is that eventually a very precise geometric understanding of HMMs can be attained that provides insight into these central problems. This approach has two main branches of historical lineage: that of hidden Markov models, and that of algebraic statistics.

Hidden Markov models were developed as statistical models in a series of papers by Leonard E. Baum and others beginning with Baum and Petrie [1966], after the description by Stratonovich [1960] of the “forward-backward” algorithm that would be used for HMM parameter estimation. HMMs have been used extensively in natural language processing and speech recognition since the development of DRAGON by Baker [1975]. As well, since

*This research was supported by the DARPA Deep Learning program (FA8650-10-C-7020)

Krogh, Mian, and Haussler [1994] used HMM for gene finding in the DNA of in *E. coli* bacteria, they have had many applications in genomics and biological sequence alignment; see also [Yoon, 2009]. Now, HMM parameter estimation is built into the measurement of so many kinds of time-series data that it would be gratuitous to enumerate them. However, the methods of algebraic statistics are not so old, and the algebraic geometry of these models is far from fully explored. They are hence an important early example for the theory to investigate.

Algebraic statistics is the application of commutative algebra and algebraic geometry to the study of statistical models, especially those models involving non-linear relations between parameters and observables. It was first described at length in the monograph *Algebraic Statistics* by Pistone, Riccomagno, and Wynn [2001]¹. Subsequent introductions to the subject include *Algebraic Statistics for Computation Biology* by Pachter and Sturmfels [2005], and *Lectures in Algebraic Statistics* by Drton, Sturmfels, and Sullivant [2009]. Also notable is *Algebraic Geometry and Statistical Learning Theory* by Watanabe [2009], for its focus on the problem of model selection from data.

To the problem of *model selection*, the algebraic analogue is *implicitization*, i.e., finding polynomial defining equations for the Zariski closures of binary hidden Markov models. Such polynomials are called *invariants* of the model: if a polynomial f is equal to a constant c at every point on the model (i.e. does not vary with the model parameters), then we encode this equation by calling $f - c$ an *invariant*. Model selection and implicitization are more than simply analogous; polynomial invariants have been used successfully in model selection by Casanellas and Fernandez-Sanchez [2006] and Eriksson [2008] for phylogenetic trees.

Invariants have been difficult to classify for hidden Markov models, perhaps due to the high codimension of the models. Bray and Morton [2005] found many invariants using linear algebra, but did not exhibit any generating sets of invariants, and in fact their search was actually for invariants of a model that was slightly modified from the HMM proper. Schönhuth [2011] found a large family of invariants arising as minors of certain non-abelian Hankel matrices, and was able to verify that such invariants generate the ideal of the 3-node binary HMM, the simplest non-degenerate HMM. However, this seemed not to be the case for models with $n \geq 4$ nodes: Schönhuth reported on a computation of J. Hauenstein which verified numerically that the 4-node model was not cut out by the Hankel minors.

In Section 3, we will make use of moment and cumulant coordinates as exposted in [Sturmfels and Zwiernik, 2011], as well as a new coordinate system on the parameter space, to find explicit defining equations for the 4-node binary HMM. The shortest quadric and cubic equations are fairly simple; to give the reader a visual sense, they look like this:

$$g_{2,1} = m_{23}m_{13} - m_2m_{134} - m_{13}m_{12} + m_1m_{124}$$

$$g_{3,1} = m_{12}^3 - 2m_1m_{12}m_{123} + m_\emptyset m_{123}^2 + m_1^2 m_{1234} - m_\emptyset m_{12}m_{1234}$$

Here each m is a moment of the observed probability distribution. These equations are not generated by Schönhuth's Hankel minors, and so provide a finer test for membership to any binary HMM of length $n \geq 4$ after marginalizing to any 4 consecutive nodes.

To the problem of *parameter identifiability*, the algebraic analogue is the generic or global injectivity or finiteness of a map of varieties that parametrizes the model, or in the case of identifying a single parameter, constancy of the parameter on the fibers of the parameterization. Sullivant et al. [2010] provide an excellent discussion of this topic in the context of identifying causal effects; see also [Meshkat, Eisenberg, and DiStefano, 2009] for a striking application to identification for ODE models in the biosciences.

In Section 4, for the purpose of parameter identification in binary hidden Markov models,

¹Pistone et al. attribute their interest in the subject to a seminar paper of Diaconis and Sturmfels [1998] circulated as a manuscript in 1993, which employed Gröbner bases to construct Markov random walks.

we express the parametrization of a binary HMM as the composition of an invertible linear map χ^{-1} , a dominant and generically finite monomial map \mathbf{q} , and a birationally invertible map ψ . An explicit inverse to the birational map is given, which allows for the easy recovery of hidden parameters in terms of observables. The components of the monomial map are *identifiable combinations* in the sense of Meshkat et al. [2009]. The formulae for recovering the hidden parameters are fairly simple when exhibited in a particular order, corresponding to a particular triangular set of generators in a union of lexicographic Gröbner bases for the model ideal. To show their simplicity, the most complicated recovery formula looks like this:

$$u = \frac{m_1 m_3 - m_2^2 + m_{23} - m_{12}}{2(m_3 - m_2)}$$

As a corollary, in Section 4.3 we find that the fibers of ϕ_n are generically zero-dimensional, consisting of two points which are equivalent under a “hidden label swapping” operation.

Section 5 describes how every BHMM lives inside a particular 9-dimensional variety called a *trace variety*, which is a GIT quotient of the space of triples of 2×2 matrices under a simultaneous conjugation action by SL_2 . As a quotient, the trace variety is not defined inside any particular ambient space. However, its coordinate ring, a *trace algebra*, was found by Sibirskii [1968] to be generated by 10 elements, which means we can embed the trace variety, and hence *all BHMMs simultaneously*, in \mathbb{C}^{10} . We prove the main results of Section 4 in the coordinates of this embedding. As a byproduct of this approach, in section Section 5.6 we find that the Zariski closures of all BHMMs with $n \geq 3$ are birational to each other.

Finally, Section 6 explores some applications of our results, including model membership testing, classification of identifiable parameters, a new grading on binary HMMs that can be used to find low-degree invariants, the geometry of equilibrium binary HMMs, and HMMs with more than two visible states.

I would like to thank my advisor, Bernd Sturmfels, and postdoctoral mentor, Shaowei Lin, for many helpful conversations and editorial suggestions on this paper.

2 Definitions

In this paper, we will be referring to binary hidden Markov *processes*, *distributions*, *maps*, *models*, *varieties*, and *ideals*. Each of these terms is used with a distinct meaning, and effort is made to keep their usages consistent and separate.

2.1 Binary Hidden Markov processes and distributions

A binary hidden Markov process is a statistical process which generates random binary sequences. It is based on the simpler notion of a binary (and not hidden) Markov *chain* process.

Definition 2.1. A **Binary Hidden Markov process** will comprise 5 data: π , T , E , and (H_t, V_t) . The pair (H_t, V_t) denotes a jointly random sequence $(H_1, V_1, H_2, V_2, \dots)$ of binary variables, also respectively called *hidden nodes* and *visible nodes*, with range $\{0, 1\}$. Often a bound n on the (discrete) time index t is also given. The joint distribution of the nodes is specified by the following:

- A row vector $\pi = (\pi_0, \pi_1)$, called the *initial distribution*, which specifies a probability distribution on the first hidden node H_1 by $\Pr(H_1 = i) = \pi_i$;

- A matrix $T = \begin{bmatrix} T_{00} & T_{01} \\ T_{10} & T_{11} \end{bmatrix}$, called the *transition matrix*, which specifies conditional “transition” probabilities by the formula $\Pr(H_t = j | H_{t-1} = i) = T_{ij}$, read as the probability of “transitioning from hidden state i to hidden state j ”.²
- A matrix $E = \begin{bmatrix} E_{00} & E_{01} \\ E_{10} & E_{11} \end{bmatrix}$, called the *emission matrix*, which specifies conditional “emission” probabilities by the formula $\Pr(V_t = j | H_t = i) = E_{ij}$, read as the probability that “hidden state i emits the visible state j ”.

To be precise, the parameter vector $\theta = (\pi, T, E)$ determines a probability distribution on the set of sequences of pairs $((H_1, V_1) \dots (H_n, V_n)) \in (\{0, 1\}^2)^n$, or if no bound n is specified, a compatible sequence of such distributions as n grows. In applications, only the joint distribution on the visible nodes $(V_1, \dots, V_n) \in \{0, 1\}^n$ is observed, and is called the *observed distribution*. This distribution is given by marginalizing (summing) over the possible hidden states of a BHM process:

$$\begin{aligned} \Pr(V = v | \theta = (\pi, T, E)) &= \sum_{h \in \{0,1\}^n} \Pr(h, v | \pi, T, E) = \sum_{h \in \{0,1\}^n} \Pr(h | \pi, T) \Pr(v | h, E) \\ &= \sum_{h \in \{0,1\}^n} \pi_{h_1} E_{h_1, v_1} \prod_{i=2}^n T_{h_{i-1} h_i} E_{h_i, v_i} \end{aligned} \quad (1)$$

Definition 2.2. A **Binary Hidden Markov *distribution*** is a probability distribution on sequences $v \in \{0, 1\}^n$ of jointly random binary variables (V_1, \dots, V_n) which arises as the observed distribution of *some* BHM *process* according to (1).

As we will see in Section 4.1, different processes (π, T, E, H_t, V_t) can give rise to the same observed distribution on the V_t , for example by permuting the labels of the hidden variables, or by other relations among the parameters.

Those already familiar with Markov models in some form may note that:

- The data (π, T, H_t) alone specify what is ordinarily called a binary Markov *chain* process on the nodes H_t . In the applications we have in mind, these nodes are unobserved variables.
- The matrices T and E are assumed to be *stationary*, meaning that they are not allowed to vary with the “time index” t of (H_t, V_t) .
- The distribution π *is not* assumed to be *at equilibrium*, i.e. we *do not* assume that $\pi T = \pi$. This allows for more diverse applications.

N.B. 2.3. The term “stationary” is sometimes also used for a process that is at equilibrium; we will reserve the term “stationary” for the constancy of matrices T, E over time.

2.2 Binary Hidden Markov maps, models, varieties, and ideals

Statistical processes come in families defined by allowing their parameters to vary, and in short, the set of probability distributions that can arise from the processes in a given family is called a *statistical model*. The Zariski closure of such a model in an appropriate complex space is an algebraic variety, and the geometry of this variety carries information about the purely algebraic properties of the model.

²[Schönhuth, 2011] uses T for different matrices, which I will later denote by P .

In a binary hidden Markov process, each row of π , T , and E must consist of non-negative reals which sum to 1, since these rows are probability distributions. Taking the complex Zariski closure of the classically compact set of such parameter vectors

$$\theta = (\pi, T, E)$$

is simply a matter of allowing the matrices (π, E, T) to have complex entries. So, for the resulting parameter ring and affine space, we write

$$\begin{aligned} \mathbb{C}[\theta] &:= \mathbb{C}[\pi_j, T_{ij}, E_{ij}] / \left(1 = \sum_j \pi_j = \sum_j T_{ij} = \sum_j E_{ij} \text{ for } i = 0, 1 \right) \\ \mathbb{C}_\theta^5 &:= \text{Spec } \mathbb{C}[\theta] \end{aligned}$$

Here Spec denotes the spectrum of a ring; see [Cox, Little, and O’Shea, 2007] for this and other fundamentals of algebraic geometry.

Convention 2.4. We will always denote complex vector spaces by \mathbb{C} to some power, with a subscript indicating the coordinate functions on that space. For example, \mathbb{C}_θ^5 is 5-dimensional space of parameter vectors $\theta = (\pi, T, E)$. We will write projective spaces similarly using \mathbb{P} raised to the dimension. Rings will usually be denoted by R with some subscripts to indicate its generators.

We denote by Δ_θ^5 the 5 real-dimensional cube of *non-negative stochastic matrices* (π, T, E) , i.e., the points of \mathbb{C}_θ^5 where $\theta = (\pi, T, E)$ have non-negative real entries. Thus

$$\Delta_\theta^5 \subseteq \mathbb{C}_\theta^5$$

Now we fix a length $|v| = n$ for our binary sequences v , and write

$$\begin{aligned} R_{p,n} &:= \mathbb{C}[p_v \mid v \in \{0, 1\}^n] & \mathbb{C}_p^{2^n} &:= \text{Spec}(R_{p,n}) \\ \bar{R}_{p,n} &:= R_{p,n} / \left(1 - \sum_{|v|=n} p_v \right) & \mathbb{C}_p^{2^n-1} &:= \text{Spec}(\bar{R}_{p,n}) \\ & & \mathbb{P}_p^{2^n-1} &:= \text{Proj}(R_{p,n}) \end{aligned}$$

We will often have occasion to consider the natural inclusions,

$$\iota_n : \mathbb{C}_p^{2^n-1} \hookrightarrow \mathbb{C}_p^{2^n} \qquad \bar{\iota}_n : \mathbb{C}_p^{2^n-1} \hookrightarrow \mathbb{P}_p^{2^n-1}$$

Definition 2.5. For $n \geq 3$,

- The **Binary Hidden Markov map** or *modeling map* on n nodes is the map $\phi_{\mathbf{BHM}(n)}$, or simply ϕ_n , given by given by (1), i.e.

$$\begin{aligned} \phi_n &: \mathbb{C}_\theta^5 \rightarrow \mathbb{C}_p^{2^n-1}, \\ \phi_n^\#(p_v) &:= \sum_{h \in \{0,1\}^n} \pi_{h_1} E_{h_1, v_1} \prod_{i=2}^n T_{h_{i-1} h_i} E_{h_i, v_i} \end{aligned}$$

The word “model” is also frequently used for the map ϕ_n . This is a very reasonable usage of the term, but I reserve “model” for the image of the allowed parameter values:

- $\mathbf{M}_{\mathbf{BHM}(n)}$, the **Binary Hidden Markov model** on n nodes, is the image

$$\bar{\iota}_n \phi_n (\Delta_\theta^5) \subseteq \mathbb{P}_p^{2^n-1},$$

i.e., the set of observed distributions which can arise from *some* BHM process, considered as a subset of $\mathbb{P}_p^{2^n-1}$ via $\bar{\iota}_n$. Being the continuous image of a classically compact set, this set is classically compact and hence classically closed.

- $\mathbf{V}_{\text{BHM}(n)}$, the **Binary Hidden Markov variety** on n nodes, is the Zariski closure of $\mathbf{M}_{\text{BHM}(n)}$ in $\mathbb{P}_p^{2^n-1}$.
- $\mathbf{I}_{\text{BHM}(n)}$, the **Binary Hidden Markov ideal** on n nodes, is the set of homogeneous polynomials which vanish on $\mathbf{M}_{\text{BHM}(n)}$, i.e., the homogeneous defining ideal of $\mathbf{V}_{\text{BHM}(n)}$. Elements of $\mathbf{I}_{\text{BHM}(n)}$ are called *invariants* of the model.

In summary, probability distributions arise from processes according to modeling maps, models are families of distributions arising from processes of a certain type, and the Zariski closure of each model is a variety whose geometry reflects the algebraic properties of the model. The ideal of the model is the same as the ideal of the variety: the definition of Zariski closure is the largest set which has the same ideal of vanishing polynomials as the model. In a rigorous sense (namely, the anti-equivalence of the categories of affine schemes and rings), the variety encodes information about the “purely algebraic” properties of the model, i.e. properties that can be stated by the vanishing of polynomials.

The number of polynomials that vanish on any given set is infinite, but by the Hilbert Basis theorem, one can always find finitely many polynomials whose vanishing implies the vanishing of all the others. This is called a *generating set* for the ideal. To compute a generating set for $\mathbf{I}_{\text{BHM}(n)}$, we will need the following proposition:

Proposition 2.6. *The ideal $\mathbf{I}_{\text{BHM}(n)}$ is the homogenization of $\ker(\phi_n^\# \circ \iota_n^\#)$ with respect to $p_\Sigma := \sum_{|v|=n} p_v$*

Proof. The affine ideal $\ker(\phi_n^\# \circ \iota_n^\#)$ cuts out the Zariski closure X of $\iota_n \circ \phi_n(\mathbb{C}_\theta^5)$ in $\mathbb{C}_p^{2^n}$, and this closure lies in the hyperplane $\{p_\Sigma = 1\} = \mathbb{C}_p^{2^n-1}$. Let X' be the projective closure of X in $\mathbb{P}_p^{2^n-1}$, so that $I(X')$ is the homogenization of $\ker(\phi_n^\# \circ \iota_n^\#)$ with respect to p_Σ .

The cube Δ_θ^5 is Zariski dense in \mathbb{C}_θ^5 , so $\iota_n \circ \phi_n(\Delta_\theta^5)$ is Zariski dense in $\iota_n \circ \phi_n(\mathbb{C}_\theta^5)$, which is Zariski dense in X , which is Zariski dense in X' . Therefore $X' = \mathbf{V}_{\text{BHM}(n)}$, and $I(X') = \mathbf{I}_{\text{BHM}(n)}$, as required. \square

3 Defining equations of $\mathbf{V}_{\text{BHM}(3)}$ and $\mathbf{V}_{\text{BHM}(4)}$

Theorem 3.1. *The homogeneous ideal $\mathbf{I}_{\text{BHM}(4)}$ of the binary hidden Markov variety $\mathbf{V}_{\text{BHM}(4)}$ is minimally generated by 21 homogeneous quadrics and 29 homogeneous cubics.*

Our fastest derivation of this result in Macaulay2 [Grayson and Stillman] uses the birational parametrization of Section 4, but in only a single step, so we defer the lengthier discussion of the parametrization until then. Modulo this dependency, the proof is described in Section 3.3, using moment coordinates (Section 3.1) and cumulant coordinates (Section 3.2).

In probability coordinates, the generators found for $\mathbf{I}_{\text{BHM}(4)}$ had the following sizes:

- Quadrics $g_{2,1}, \dots, g_{2,21}$: respectively 8, 8, 12, 14, 16, 21, 24, 24, 26, 26, 28, 32, 32, 41, 42, 43, 43, 44, 45, 72, 72 probability terms.
- Cubics $g_{3,1}, \dots, g_{3,29}$: respectively 32, 43, 44, 44, 44, 52, 52, 56, 56, 61, 69, 71, 74, 76, 78, 81, 99, 104, 109, 119, 128, 132, 148, 157, 176, 207, 224, 236, 429 probability terms.

As a motivation for introducing moment coordinates, we note here that these generators have considerably fewer terms when written in terms of moments:

- Quadrics $g_{2,1}, \dots, g_{2,21}$: respectively 4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 6, 6, 8, 8, 8, 8, 8, 10, 10, 10, 17 moment terms.

- Cubics $g_{3,1}, \dots, g_{3,29}$: respectively 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 8, 8, 8, 8, 10, 10, 10, 10, 10, 12, 12, 13, 14, 16, 18, 21, 27, 35 moment terms.

To give a sense of how these polynomials look in moment coordinates, the shortest quadric and cubic are

- $g_{2,1} = m_{23}m_{13} - m_2m_{134} - m_{13}m_{12} + m_1m_{124}$, and
- $g_{3,1} = m_{12}^3 - 2m_1m_{12}m_{123} + m_\emptyset m_{123}^2 + m_1^2 m_{1234} - m_\emptyset m_{12}m_{1234}$.

Let us compare this ideal with $\mathbf{I}_{\mathbf{BHM}(3)}$, the homogeneous defining ideal of $\mathbf{V}_{\mathbf{BHM}(3)}$. Schönhuth [2011] found that $\mathbf{I}_{\mathbf{BHM}(3)}$ is precisely the ideal of 3×3 minors of the following matrix:

$$A_{3,3} = \begin{bmatrix} p_{000} + p_{001} & p_{000} & p_{100} \\ p_{010} + p_{011} & p_{001} & p_{101} \\ p_{100} + p_{101} & p_{010} & p_{110} \\ p_{110} + p_{111} & p_{011} & p_{111} \end{bmatrix} \quad (2)$$

Schönhuth defines an analogous matrix $A_{n,3}$ for $\mathbf{V}_{\mathbf{BHM}(n)}$, but then remarks that J. Hauenstein has found, using numerical rank deficiency testing [Bates, Hauenstein, Peterson, and Sommese, 2010] with the algebraic geometry package Bertini [Bates, Hauenstein, Sommese, and Wampler], that $\text{minors}_3(A_{n,3})$ does not cut out $\mathbf{V}_{\mathbf{BHM}(n)}$ when $n = 4$. In general, Schönhuth shows that $\mathbf{I}_{\mathbf{BHM}(n)} = (\text{minors}_3(A_{n,3}) : \text{minors}_2(B_{n,2}))$ for a particular 2×3 matrix $B_{n,2}$, but computing generators for this colon ideal is a costly operation, and so no generating set for $\mathbf{I}_{\mathbf{BHM}(n)}$ was found for any $n \geq 4$ by this method. Instead, here we will make use of *moment coordinates* and *cumulant coordinates* as exposted in [Sturmfels and Zwiernik, 2011].

3.1 Moment coordinates

Moments are particular linear expressions in probabilities. They can be derived from a moment generating function as in [Sturmfels and Zwiernik, 2011], but in our case, moments can be expressed simply by the following rule: we order $\{0, 1\}^n$ by strict dominance, i.e. $v \geq w$ iff $v_i \geq w_i$ for all i , and then

$$m_v := \sum_{w \geq v} p_w \in R_{p,n} \quad (3)$$

Since all our variables are binary, with the usual algebraic statistical convention that a “+” subscript denotes an index to be summed over, we can view the conversion from moments to probabilities as “replacing zeros by + signs”. For example, $m_{10010} = p_{1++1+}$. The ring elements $m_v \in R_{p,n}$ provide alternative linear coordinates on $\mathbb{P}_p^{2^n-1}$ in which it turns out that some previously intractable BHM computations are simplified and become feasible.

For a more compact notation, a binary string v of length n is the indicator function of a unique subset I of $[n] = \{1, \dots, n\}$, so we also write m_I to represent m_v . For example, $m_{0000} = m_\emptyset$, $m_{1000} = m_1$, and $m_{0101} = m_{24}$. From (3) we can see that m_I actually represents a marginal probability: $m_I = \Pr(V_i = 1 \text{ for all } i \in I)$. Thus, in the context of BHMMS, no confusion results if we write m_I without specifying the value of n . To be precise, if $I \subseteq [n]$ and I' denotes I considered as a subset of $[n']$ for some $n' > n$, then

$$\phi_n^\#(m_I) = \phi_{n'}^\#(m_{I'}) \quad (4)$$

This can be seen in many ways, for example using the Baum formula for moments (Proposition 5.1) as explained in Section 5.3.

Just as for probabilities, for moments we define rings and spaces

$$\begin{aligned} R_{m,n} &:= \mathbb{C}[m_I \mid I \subseteq [n]] & \mathbb{C}_m^{2^n} &:= \text{Spec}(R_{m,n}) \\ \overline{R}_{m,n} &:= R_{m,n}/\langle 1 - m_\emptyset \rangle & \mathbb{C}_m^{2^n-1} &:= \text{Spec}(\overline{R}_{m,n}) \\ & & \mathbb{P}_m^{2^n-1} &:= \text{Proj}(R_{m,n}), \end{aligned} \quad (5)$$

To avoid having notation for too many ring isomorphisms, we adopt:

Convention 3.2. Using (3), we will usually treat m_I as a literal element of $R_{p,n}$, thus creating literal identifications

$$R_{m,n} = R_{p,n}, \quad \overline{R}_{m,n} = \overline{R}_{p,n}, \quad \mathbb{C}_m^{2^n} = \mathbb{C}_p^{2^n}, \quad \mathbb{P}_m^{2^n-1} = \mathbb{P}_p^{2^n-1}, \quad \text{and} \quad \mathbb{C}_m^{2^n-1} = \mathbb{C}_p^{2^n-1}. \quad (6)$$

Note that, for example, we obtain natural ring inclusions

$$R_{m,n} \subseteq R_{m,n'}$$

whenever $n < n'$, which respect the BHM maps ϕ_n by (4).

As a first application of moment coordinates, we have

Proposition 3.3. *The homogeneous ideal $\mathbf{I}_{\text{BHM}(3)}$ is generated in moment coordinates by the 3×3 minors of the matrix*

$$A'_{3,3} = \begin{bmatrix} m_{000} & m_{000} & m_{100} \\ m_{010} & m_{001} & m_{101} \\ m_{100} & m_{010} & m_{110} \\ m_{110} & m_{011} & m_{111} \end{bmatrix} = \begin{bmatrix} m_\emptyset & m_\emptyset & m_1 \\ m_2 & m_3 & m_{13} \\ m_1 & m_2 & m_{12} \\ m_{12} & m_{23} & m_{123} \end{bmatrix}$$

In particular, the projective variety $\mathbf{V}_{\text{BHM}(3)}$ is cut out by these minors.

Proof. Observe that Schönhuth's matrix $A_{3,3}$ in (2) is equivalent under elementary row/column operations to $A'_{3,3}$, so $\text{minors}_3 A'_{3,3} = \text{minors}_3 A_{3,3} = \mathbf{I}_{\text{BHM}(3)}$. \square

Proposition 3.4. *The ideal $\mathbf{I}_{\text{BHM}(n)}$ is the homogenization of $\ker(\phi_n^\#)$ with respect to m_\emptyset .*

Proof. From Proposition 2.6 we know that $\mathbf{I}_{\text{BHM}(n)}$ is the homogenization of $\ker(\phi_n^\# \circ \iota_n^\#)$ with respect to $m_\emptyset = \sum_{|v|=n} p_v$. From (5), we can identify $\overline{R}_{m,4}$ with the polynomial subring of $R_{m,4}$ obtained by omitting m_\emptyset , so that $\ker(\phi_4^\# \circ \iota_4^\#) = \ker(\phi_4^\#) + \langle 1 - m_\emptyset \rangle$. Since the additional generator $1 - m_\emptyset$ homogenizes to 0, $\ker(\phi_4^\#)$ has the same homogenization as $\ker(\phi_4^\# \circ \iota_4^\#)$, hence the result. \square

3.2 Cumulant coordinates

Cumulants are non-linear expressions in moments or probabilities which seem to allow even faster computations with binary hidden Markov models. Let

$$\begin{aligned} R_{k,n} &:= \mathbb{C}[k_I \mid I \subseteq [n]] \\ \overline{R}_{k,n} &:= R_{k,n}/\langle k_\emptyset \rangle \\ \mathbb{C}_k^{2^n-1} &:= \text{Spec}(\overline{R}_{k,n}) \end{aligned}$$

where, as with moments, we may freely alternate between writing k_v and writing k_I , where I is the set of positions where 1 occurs in v . For building generating functions, let x_1, \dots, x_n

be indeterminates, and write $x^v = x^I$ for $x_1^{v_1} \cdots x_n^{v_n} = \prod_{i \in I} x_i$. Let J be the ideal generated by all the squares x_i^2 . Following [Sturmfels and Zwiernik, 2011], we define the *moment* and *cumulant generating functions*, respectively, as

$$f_m(x) := \sum_{I \subseteq [n]} m_I x^I \in \overline{R}_{m,n}[x]/J \quad f_k(x) := \sum_{I \subseteq [n]} k_I x^I \in \overline{R}_{k,n}[x]/J$$

We now define changes of coordinates

$$\kappa_n : \mathbb{C}_m^{2^n-1} \rightarrow \mathbb{C}_k^{2^n-1} \quad \kappa_n^{-1} : \mathbb{C}_k^{2^n-1} \rightarrow \mathbb{C}_m^{2^n-1}$$

by the formulae

$$\begin{aligned} \kappa_n^\#(f_k) &= \log(f_m) = \frac{(f_m - 1)}{1} + \cdots + (-1)^{n+1} \frac{(f_m - 1)^n}{n} \\ \kappa_n^{-\#}(f_m) &= \exp(f_k) = 1 + \frac{(f_k)}{1} + \cdots + \frac{(f_k)^n}{n!} \end{aligned} \quad (7)$$

That is, we let $\kappa_n^\#(k_I)$ be the coefficient of x^I in the Taylor expansion of $\log f_m$ about 1, and let $\kappa_n^{-\#}(m_I)$ be the coefficient of x^I in the Taylor expansion of $\exp f_k$ about 0. Note that in the relevant coordinate rings $\overline{R}_{m,n}$ and $\overline{R}_{k,n}$, $m_\emptyset = 1$ and $k_\emptyset = 0$. This is why we only need to compute the first n terms of each Talyor expansion: the higher terms all vanish modulo the ideal J .

Proposition 3.5. *The expressions $\kappa_n^\#(k_I)$ and $\kappa_n^{-\#}(m_I)$, i.e. writing of cumulants in terms of moments and conversely, do not depend on n .*

Proof. In [Sturmfels and Zwiernik, 2011], these formulae are re-expressed using Möbius functions, which do not depend on the generating function description above, and in particular do not depend on n . \square

3.3 Deriving $\mathbf{I}_{\mathbf{BHM}(4)}$ in Macaulay2

This section describes the proof of Theorem 3.1 using Macaulay2. These computations were carried out on a Toshiba Satellite P500 laptop running Ubuntu 10.04, with an Intel Core i7 Q740 .73 GHz CPU and 8gb of RAM. In light of Proposition 3.4, we will aim to compute $\ker(\phi_4^\# \circ \iota_4^\#)$, which can be understood geometrically as the (non-homogeneous) ideal of the standard affine patch of $\mathbf{V}_{\mathbf{BHM}(4)}$ where $m_\emptyset = \sum_{|v|=4} p_v = 1$. To reduce the number of variables, as in Proposition 3.4 we continue to make the identification

$$\overline{R}_{m,4} = \mathbb{C}[m_I | \emptyset \neq I \subseteq [4]] \subseteq R_{m,4}$$

We begin by providing Macaulay2 with the map $\phi_4^\# : \overline{R}_{m,4} \rightarrow \mathbb{C}[\theta]$ in moment coordinates (Section 3.1), because probability coordinates result in longer, higher degree expressions. This can be done by composing the expression of $\phi_n^\#(p_v)$ in Definition 2.5 with the expression of $m_v = m_I$ in (3), or alternatively using the Baum formula for moments (Proposition 5.1), which involves many fewer arithmetic operations.

Macaulay2 runs out of memory (8gb) trying to compute $\ker(\phi_4^\#)$, and as expected, this memory runs out even sooner in probability coordinates, so we use cumulant coordinates instead (Section 3.2). We input

$$\kappa_4^\# : \overline{R}_{k,4} \rightarrow \overline{R}_{m,4}$$

using coefficient extraction from (7), and compute the composition $\phi_4^\# \circ \kappa_4^\#$. Then, it is possible to compute

$$\mathbf{I}_{k,4} := \ker(\phi_4^\# \circ \kappa_4^\#)$$

which takes **around 1.5 hours**. Alternatively, we can compute $\mathbf{I}_{k,4}$ using the birational parameterization $\bar{\psi}_4$ of Section 4 in place of ϕ_4 , which takes **less than 1 second** and yields 100 generators for $\mathbf{I}_{k,4}$.

Subsequent computations run out of memory with this set of 100 generators, so we must take some steps to simplify it. Macaulay2's **trim** command reduces the number of generators of $\mathbf{I}_{k,4}$ to 46 in **under 1 second**. We then order these 46 generators lexicographically, first by degree and then by number of terms, and eliminate redundant generators in reverse order, which takes **19 seconds**. The result is an inclusion-minimal, non-homogeneous generating set for $\mathbf{I}_{k,4}$ with 35 generators: 24 quadrics and 11 cubics.

Now we compute $\mathbf{I}_{m,4} := \kappa^\#(\mathbf{I}_{k,4}) = \kappa^\#(\ker(\phi_4^\# \circ \kappa_4^\#)) = \ker(\phi_4^\#)$, i.e., we push forward the 35 generators for $\mathbf{I}_{k,4}$ under the non-linear ring isomorphism $\kappa_4^\#$ to obtain 35 generators for $\mathbf{I}_{m,4} = \ker(\phi_4^\#)$: 2 quadrics, 7 cubics, 16 quartics, 5 quintics, and 5 sextics. In **under 1 second**, Macaulay2's **trim** command computes a new set of 39 generators for $\mathbf{I}_{m,4}$ with lower degrees: 21 quadrics, 14 cubics, and 4 quartics, which turns out to **save around 1 hour** of computing time in what follows. These generators have many terms each, and eliminating redundant generators as in the previous paragraph turns out to be too slow to be worth it here, taking more than 2 hours, so we omit this step.

Finally, we apply Proposition 3.4 to compute $\mathbf{I}_{\text{BHM}(4)}$ as the homogenization of $\mathbf{I}_{m,4}$ with respect to m_\emptyset . In Macaulay2, this is achieved by homogenizing the 39 generators for $\mathbf{I}_{m,4}$ with respect to m_\emptyset and then saturating the ideal they generate with respect to m_\emptyset . This saturation operation takes about **29 minutes**, and yields a minimal generating set of 50 polynomials: 21 quadrics and 29 cubics. Since probabilities are linear in moments, their degrees are the same in probability coordinates. Moreover, since these are homogeneous generators for a homogeneous ideal, they are minimal in a very strong sense:

Corollary 3.6. *Any inclusion-minimal homogeneous generating set for $\mathbf{I}_{\text{BHM}(4)}$ in probability or moment coordinates must contain exactly 21 quadrics and 29 cubics.*

We still do not know a generating set for $\mathbf{I}_{\text{BHM}(5)}$. Macaulay2 runs out of memory (8gb) attempting to compute $\mathbf{I}_{k,5}$, even using the birational parametrization of Section 4. The author has also attempted this computation using the *tree cumulants* of Smith and Zwiernik [2010] in place of cumulants, but again Macaulay2 runs out of memory trying to compute the first kernel. Presumably the subsequent saturation step would be even more difficult.

4 Birational parametrization of BHMMs

Theorem 4.1 (Birational Parameter Theorem). *There is a generically two-to-one, dominant morphism $\mathbb{C}_\theta^5 \rightarrow \mathbb{C}^5$ such that, for each $n \geq 3$, the binary hidden Markov map ϕ_n factors uniquely as*

$$\begin{array}{ccccc} \mathbb{C}_\theta^5 & \longrightarrow & \mathbb{C}^5 & \xrightarrow{\bar{\psi}_n} & \mathbb{C}_p^{2^n-1} \\ & \searrow & & \nearrow & \\ & & & \phi_n & \end{array}$$

and each $\bar{\psi}_n : \mathbb{C}^5 \rightarrow \mathbf{V}_{\text{BHM}(n)}$ has a birational inverse map ρ_n :

$$\begin{array}{ccc}
& \overline{\psi}_n & \\
& \curvearrowright & \\
\mathbb{C}^5 & & \mathbf{V}_{\mathbf{BHM}(\mathbf{n})} \\
& \curvearrowleft & \\
& \rho_n &
\end{array}$$

In particular, $\mathbf{V}_{\mathbf{BHM}(\mathbf{n})}$ is always a rational projective variety, i.e., birationally equivalent to \mathbb{P}^5 .

This theorem will be proven in Section 5.6 using trace algebras and the Baum formula for moments. In the course of this section and Section 5 we will exhibit formulae for $\overline{\psi}_n$ and their inverses ρ_n . The inverse map ρ_3 has a number of practical uses, to be explored in Section 6.

Our first step toward Theorem 4.1 is to re-parametrize \mathbb{C}_θ^5 .

4.1 A linear reparametrization of \mathbb{C}_θ^5

Since the hidden variables H_t are never observed, there is no change in the final expression of p_v in Definition 2.5 if we swap the labels $\{0, 1\}$ of all the H_t simultaneously. This swapping is equivalent to an action of the elementary permutation matrix $\sigma = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$:

$$\begin{aligned}
\mathbf{sw} : \mathbb{C}_\theta^5 &\rightarrow \mathbb{C}_\theta^5 \\
\theta = (\pi, T, E) &\mapsto (\pi\sigma, \sigma^{-1}T\sigma, \sigma^{-1}E)
\end{aligned} \tag{8}$$

(In our case $\sigma^{-1} = \sigma$, but the form above generalizes to permutations of larger hidden alphabets.) Hence we have that $\Pr(v \mid \pi, T, E) = \Pr(v \mid \mathbf{sw}(\pi, T, E))$, i.e. $\phi_n = \phi_n \circ \mathbf{sw}$.

We will make essential use of a linear parametrization of \mathbb{C}_θ^5 in which \mathbf{sw} has a simple form. Our new parameter vector will be $\eta_0 := (a_0, b, c_0, u, v_0)$, with subscript 0's to be explained shortly. Although we have already used the letter v at times to represent visible binary strings, we hope that the context will be clear enough to avoid confusion between these usages. We let

$$\begin{aligned}
\mathbb{C}[\eta_0] &:= \mathbb{C}[a_0, b, c_0, u, v_0], & \mathbb{C}_{\eta_0}^5 &:= \text{Spec}(\mathbb{C}[\eta_0]), \\
\pi &= \frac{1}{2} [1 - a_0, & 1 + a_0] \\
T &= \frac{1}{2} \begin{bmatrix} 1 + b - c_0, & 1 - b + c_0 \\ 1 - b - c_0, & 1 + b + c_0 \end{bmatrix} & E &= \begin{bmatrix} 1 - u + v_0, & u - v_0 \\ 1 - u - v_0, & u + v_0 \end{bmatrix}
\end{aligned} \tag{9}$$

(The rightmost column of E is made intentionally homogeneous in the parameters.) This defines a linear isomorphism $\chi : \mathbb{C}_{\eta_0}^5 \xrightarrow{\sim} \mathbb{C}_\theta^5$. For convenience, we write \mathbf{sw} also for $\chi \circ \mathbf{sw} \circ \chi^{-1}$. In these coordinates, \mathbf{sw} acts by

$$a_0 \mapsto -a_0, \quad b \mapsto b, \quad c_0 \mapsto -c_0, \quad u \mapsto u, \quad v_0 \mapsto -v_0$$

In other words, swapping the signs of the subscripted variables a_0, c_0, v_0 has the same effect as acting on the matrices π, T, E by σ as in (8), i.e., relabeling the hidden alphabet.

We will write $\Delta_{\eta_0}^5 := \chi^{-1}\Delta_\theta^5$, the classically compact set of real points giving rise to non-negative stochastic matrices. Let us write ψ_n for the new parametrization $\phi_n \circ \chi$:

$$\psi_n : \mathbb{C}_{\eta_0}^5 \xrightarrow{\chi} \mathbb{C}_\theta^5 \xrightarrow{\phi_n} \mathbb{C}_p^{2^n-1}$$

Recall that we write $\iota_n : \mathbb{C}_p^{2^n-1} \hookrightarrow \mathbb{C}_p^{2^n}$ and $\bar{\iota}_n : \mathbb{C}_p^{2^n-1} \hookrightarrow \mathbb{P}_p^{2^n-1}$ for the natural inclusions, so that now $\mathbf{V}_{\mathbf{BHM}(\mathbf{n})} = \overline{\text{image}(\bar{\iota}_n \psi_n)}$. The morphism ψ_n defines a ring map $\psi_n^\# : \overline{R}_{p,n} \rightarrow \mathbb{C}[\eta_0]$, and $\ker \psi_n^\# = \ker \phi_n^\#$ cuts out $\mathbf{V}_{\mathbf{BHM}(\mathbf{n})}$ in $\text{image}(\bar{\iota}_n)$, i.e., the standard affine patch where $\sum_{|v|=n} p_v = 1$.

4.2 Introducing the birational parameters

Since $\psi_n \circ \mathbf{sw} = \psi_n$, by classical invariant theory the ring map $\psi_n^\# : \overline{R}_{p,n} \rightarrow \mathbb{C}[\eta_0]$ must land in the subring of invariants $\mathbb{C}[\eta_0]^{\mathbf{sw}} = \mathbb{C}[b, u, a_0^2, c_0^2, v_0^2, a_0c_0, a_0v_0, c_0v_0]$. However, $\psi_n^\#$ in fact factors through a smaller subring, conveniently generated by 5 elements:

Lemma 4.2 (Parameter Subring Lemma). *For all $n \geq 3$, the ring map $\psi_n^\#$ lands in the subring*

$$\mathbb{C}[\eta] := \mathbb{C}[a, b, c, u, v]$$

of $\mathbb{C}[\eta_0]$, where $a = a_0v_0$, $c = c_0v_0$, $v = v_0^2$.

The proof of this key lemma will be given in Section 5.5 after introducing trace algebras. To interpret its geometric consequences, write $\mathfrak{q}^\#$ for the subring inclusion

$$\mathfrak{q}^\# : \mathbb{C}[\eta] \hookrightarrow \mathbb{C}[\eta_0]$$

$$a \mapsto a_0v_0, \quad b \mapsto b, \quad c \mapsto c_0v_0, \quad u \mapsto u, \quad v \mapsto v_0^2,$$

write $\overline{\psi}_n^\# : \overline{R}_{p,n} \rightarrow \mathbb{C}[\eta]$ for the factorization of $\psi_n^\#$ through $\mathfrak{q}^\#$, and write $\mathbb{C}_\eta^5 := \text{Spec } \mathbb{C}[\eta]$. The result:

Corollary 4.3. *The following diagram of dominant maps commutes:*

$$\begin{array}{ccccccc} \mathbb{C}_\theta^5 & \xrightarrow{\chi^{-1}} & \mathbb{C}_{\eta_0}^5 & \xrightarrow{\mathfrak{q}} & \mathbb{C}_\eta^5 & \xrightarrow{\overline{\psi}_n} & \mathbf{V}_{\text{BHM}(\mathbf{n})} \\ & \searrow & & & & \nearrow & \\ & & & & & & \phi_n \end{array}$$

where $\chi : \mathbb{C}_{\eta_0}^5 \rightarrow \mathbb{C}_\theta^5$ is the isomorphism given by (9), and \mathfrak{q} is generically two-to-one.

This corollary in particular implies the first part of the Birational Parameter Theorem (4.1), by taking $\mathfrak{q} \circ \chi^{-1} : \mathbb{C}_\theta^5 \rightarrow \mathbb{C}_\eta^5$ as the generically 2 : 1 map.

Remark 4.4. The map \mathfrak{q} is only dominant, and not surjective; for example, it misses the point $(1, 0, 0, 0, 0)$.

Corollary 4.5. *For all $n \geq 3$, $\mathbf{V}_{\text{BHM}(\mathbf{n})} = \overline{\text{image}}(\overline{\tau}_n \overline{\psi}_n)$.*

Proof. Since \mathfrak{q} is dominant and χ^{-1} is an isomorphism, $\overline{\text{image}}(\overline{\tau}_n \overline{\psi}_n) = \overline{\text{image}}(\overline{\tau}_n \psi_n \mathfrak{q}) = \overline{\text{image}}(\overline{\tau}_n \psi_n \mathfrak{q} \chi^{-1}) = \overline{\text{image}}(\overline{\tau}_n \phi_n) =: \mathbf{V}_{\text{BHM}(\mathbf{n})}$. \square

The unique factorization map $\overline{\psi}_n^\#$ can be computed directly in Macaulay2 for small n . The expressions in moment coordinates are simpler than in probabilities, so we present these in the following proposition.

Proposition 4.6. *The map $\overline{\psi}_3^\#$ is given in moment coordinates by*

$$\begin{aligned}
m_\emptyset &= m_{000} \mapsto 1 \\
m_1 &= m_{100} \mapsto a + u \\
m_2 &= m_{010} \mapsto ab + c + u \\
m_3 &= m_{001} \mapsto ab^2 + bc + c + u \\
m_{12} &= m_{110} \mapsto abu + ac + au + cu + u^2 + bv \\
m_{13} &= m_{101} \mapsto ab^2u + abc + bcu + b^2v + ac + au + cu + u^2 \\
m_{23} &= m_{011} \mapsto ab^2u + abc + abu + bcu + c^2 + 2cu + u^2 + bv \\
m_{123} &= m_{111} \mapsto ab^2u^2 + 2abcu + abu^2 + bcu^2 + b^2uv + ac^2 + 2acu \\
&\quad + c^2u + au^2 + 2cu^2 + u^3 + abv + bcv + 2bu v
\end{aligned}$$

We will eventually prove the Birational Parameter Theorem (4.1) by marginalization to the case $n = 3$, which we can prove here:

Proposition 4.7. *The following triangular set of equations hold on the graph of $\overline{\psi}_3$, after clearing denominators, and can thus be used to recover parameters from observed moments where the denominators are non-zero:*

$$\begin{aligned}
b &= \frac{m_3 - m_2}{m_2 - m_1} \\
u &= \frac{m_1 m_3 - m_2^2 + m_{23} - m_{12}}{2(m_3 - m_2)} \\
a &= m_1 - u \\
c &= a - ba + m_2 - m_1 \\
v &= a^2 - \frac{m_1 m_2 - m_{12}}{b}
\end{aligned}$$

(This result actually holds for all ϕ_n with $n \geq 3$, because of Proposition 5.2.)

Proof. These equations can be checked with direct substitution by hand from Proposition 4.6. Regarding the derivation, they can be obtained in Macaulay2 by computing two Gröbner bases of the elimination ideal $I = \langle m_v - \psi_3(m_v) \mid v \in \{0, 1\}^3 \rangle$ over the ring $\mathbb{C}_m^{2^3}$, in Lex monomial order: once in the ring $R_{m,3}[v, c, a, b, u]$, and once in $R_{m,3}[v, c, u, b, a]$. Each variable occurs in the leading term of a some generator in one of these two bases with a simple expression in moments as its leading coefficient. We solve each such generator (set to 0) for the desired parameter. \square

Corollary 4.8. *The map $\overline{\psi}_3 : \mathbb{C}^5 \rightarrow \mathbf{V}_{\text{BHM}(3)}$ has a birational inverse ρ_3 . The map $\rho_3^\#$ on moment coordinate functions is given by:*

$$\begin{aligned}
a &\mapsto \frac{m_2^2 + m_3 m_1 - 2m_2 m_1 - m_{23} + m_{12}}{2(m_3 - m_2)} & u &\mapsto \frac{-m_2^2 + m_3 m_1 + m_{23} - m_{12}}{2(m_3 - m_2)} \\
b &\mapsto \frac{m_3 - m_2}{m_2 - m_1} & v &\mapsto \frac{\text{num}(v)}{4(m_3 - m_2)^2} \\
c &\mapsto \frac{\text{num}(c)}{2(m_2 - m_1)(m_3 - m_2)}, \text{ where}
\end{aligned}$$

$$\begin{aligned}
\text{num}(c) &= -m_1m_2^2 + m_1^2m_3 + m_2^2m_3 - m_1m_3^2 - m_1m_{12} \\
&\quad + 2m_2m_{12} - m_3m_{12} + m_1m_{23} - 2m_2m_{23} + m_3m_{23}, \text{ and} \\
\text{num}(v) &= m_2^4 - 2m_1m_2^2m_3 + m_1^2m_3^2 - 2m_2^2m_{12} - 2m_1m_3m_{12} + 4m_2m_3m_{12} \\
&\quad + 4m_1m_2m_{23} - 2m_2^2m_{23} - 2m_1m_3m_{23} + m_{12}^2 - 2m_{12}m_{23} + m_{23}^2.
\end{aligned}$$

(This result will also hold for all ϕ_n with $n \geq 3$ by Proposition 5.2.)

Proof. This can be derived by substituting the solutions for u , a , and b in the previous propositions into the subsequent solutions for a , c , and v . Alternatively, it can be checked by direct substitution in Macaulay2, i.e., one computes that $\overline{\psi}_3^\# \circ \rho^\#(\theta) = \theta$ for each birational parameter $\theta \in \{a, b, c, u, v\}$. \square

The expressions in Corollary 4.8 are considerably simpler in moment coordinates than in probabilities. Comparing the number of terms, the numerators for a, b, c, u, v respectively have sizes 5, 2, 10, 4, and 12 in moment coordinates, versus sizes 22, 4, 56, 22, and 190 in probability coordinates. This explains in part why Macaulay2's Gröbner basis computations execute in moment coordinates with much less time and memory.

4.3 Statistical interpretation of the birational inverse ρ_3

It turns out that the factors appearing in the denominators of Corollary 4.8 defining ρ_3 have simple factorizations in terms of the rational and birational parameters:

- $m_3 - m_2$ appears in the denominator of all $\rho_3(\theta)$ except $\rho_3(b)$, and

$$m_3 - m_2 \xrightarrow{\overline{\psi}_3} (b)(ab - a + c) \xrightarrow{q} (b)(v_0)(a_0b - a_0 + c_0)$$

- $m_2 - m_1$ appears in the denominator of $\rho_3(b)$ and $\rho_3(c)$, and

$$m_2 - m_1 \xrightarrow{\overline{\psi}_3} ab - a + c \xrightarrow{q} (v_0)(a_0b - a_0 + c_0)$$

Let us pause to reflect on the meaning of these factors.

- The factor v_0 occurs in $\det(E) = 2v_0$, hence $v = v_0^2 = 0$ iff the hidden Markov chain has “no effect” on the observed variables. The image locus $\psi_3(\{v_0 = 0\})$ can thus be modeled by a sequence of IID coin flips with distribution $E_0 = E_1 = (1 - u, u)$, so the BHMM is an unlikely model choice. This is a **one-dimensional submodel**.
- The factor b occurs in $\det(T) = b$, hence $b = 0$ iff each hidden node has “no effect” on the subsequent hidden nodes. In this case, the observed process can be modeled as a sequence of independent coin flips, the first flip having distribution πE and subsequent flips being IID having distribution $T_0E = T_1E$, a process which requires one parameter for each of these two different distributions. The image locus $\psi_3(\{b = 0\})$ is hence a **two-dimensional submodel**.
- The factor $a_0b - a_0 + c_0$ occurs in $\pi T - \pi = \frac{1}{2}(-a_0b + a_0 - c_0, a_0b - a_0 + c_0)$. Hence $a_0b - a_0 + c_0 = 0$ iff π is a fixed point of T , i.e. the hidden Markov chain is *at equilibrium*. We may define the Equilibrium Binary Hidden Markov model, EBHMM, by restricting ψ_3 to the locus $\{a_0b - a_0 + c_0 = 0\}$, which turns out to yield a **four-dimensional submodel**. The geometry of this submodel will be examined in future work in order to identify the learning coefficients of BHMM fibers.

4.4 Computational advantages of moments, cumulants, and birational parameters

Our approach has been to work with moments m_v and cumulants k_v instead of probabilities p_v , and the birational parameters a, b, c, u, v instead of the matrix entries π_1, t_{i1}, e_{i1} . Other than the theoretical advantage that the model map is generically injective on the birational parameter space, significant computation gains in Macaulay2 also result from these choices (see Section 3.3 for laptop specifications):

- Computing $\ker \bar{\psi}_3 = \ker \phi_3$, the affine defining ideal of $\mathbf{V}_{\text{BHM}(3)}$, took less than 1 second in Macaulay2 when using the birational parameters, compared to 25 seconds when using the matrix entries and moments, and 15 minutes when using the matrix entries and probabilities.
- Computing $\ker \bar{\psi}_4 = \ker \phi_4$, the affine defining ideal of $\mathbf{V}_{\text{BHM}(4)}$ took less than 1 second in Macaulay2 when using the birational parameters and *cumulant coordinates* [Sturmfels and Zwiernik, 2011], compared to 1.5 hours when using the matrix entries and cumulant coordinates, and running out of memory (8gb) when using the matrix entries and probabilities.

5 A simultaneous embedding of all BHMMs

In this section, we exhibit an embedding of every BHMM in a particular *trace variety* called $\text{Spec } C_{2,3}$, which itself can be embedded in \mathbb{C}^{10} . We use these coordinates to prove the Birational Parameter Theorem (4.1) and the Parameter Subring Lemma (4.2), which were stated without proof.

For this, we will define a map ϕ_∞ through which all the ϕ_n factor, and using a version of the Baum formula for moments, we factor this map further through $\text{Spec } C_{2,3}$. Then we use a finite set 10 of generators of the ring $C_{2,3}$ exhibited by [Sibirskii, 1968] to show that the image of $\psi_\infty^\# := \chi^\# \circ \phi_\infty^\#$ lands in the desired subring $\mathbb{C}[\eta]$, and write $\bar{\psi}_n$ for the factorization. Finally, by marginalizing to the case $n = 3$, we obtain a birational inverse for $\bar{\psi}_n$ from the map ρ_3 given in Corollary 4.8.

5.1 Marginalization maps

For each pair of integers $n' \geq n \geq 1$, the *marginalization map* $\mu_n^{n'} : \mathbb{C}_p^{2^{n'}} \rightarrow \mathbb{C}_p^{2^n}$ is given by

$$\mu_n^{n'\#}(p_v) := \sum_{|w|=n'-n} p_{vw}$$

These restrict to maps $\mu_n^{n'} : \mathbb{C}_p^{2^{n'}-1} \rightarrow \mathbb{C}_p^{2^n-1}$, and define *rational maps* $\mu_n^{n'} : \mathbb{P}_p^{2^{n'}-1} \dashrightarrow \mathbb{P}_p^{2^n-1}$.

In moment coordinates, these maps are actually coordinate projections: $\mu_n^{n'\#}(m_v) = m_{v\bar{0}}$ where $\bar{0}$ denotes a sequence of $n' - n$ zeros. In fact, using the subset notation for moments m_I , the corresponding ring maps are literal inclusions: $\mu_n^{n'\#}(m_I) = m_I$. In other words, $\mu_n^{n'} : \mathbb{C}_m^{2^{n'}} \rightarrow \mathbb{C}_m^{2^n}$ is just the map which forgets those m_I where $I \not\subseteq [n]$.

5.2 The Baum formula for moments

Equation (1) involves $O(2^n)$ addition operations. There is a faster way to compute $\phi_n^\#(p_v)$, using $O(n)$ arithmetic operations, by treating the BHM process as a *finitary process* [Schönhuth,

2011]. We define two new matrices³

$$(P_i)_{jk} := E_{ji}T_{jk} = \Pr(V_t = i \text{ and } H_{t+1} = k \mid H_t = j \text{ and } \pi, E, T), \text{ that is,}$$

$$P_0 := \begin{bmatrix} T_{00}E_{00} & T_{01}E_{00} \\ T_{10}E_{10} & T_{11}E_{10} \end{bmatrix} \quad \text{and} \quad P_1 := \begin{bmatrix} T_{00}E_{01} & T_{01}E_{01} \\ T_{10}E_{11} & T_{11}E_{11} \end{bmatrix}$$

Writing $\mathbb{1}$ for the vector $(\frac{1}{1})$ we obtain the matrix expression $\phi^\#(p_v) = \pi P_{v_1} P_{v_2} \cdots P_{v_n} \mathbb{1}$ which involves only $4n + 2$ multiplications and $2n + 1$ additions. This is known as the Baum formula. We can rewrite this formula as a trace product of 2×2 matrices:

$$\phi^\#(p_v) = \text{trace}(\pi P_{v_1} P_{v_2} \cdots P_{v_n} \mathbb{1}) = \text{trace}((\mathbb{1}\pi) P_{v_1} P_{v_2} \cdots P_{v_n})$$

To create an analogue of this formula in moment coordinates, we let

$$M_0 := P_0 + P_1 = T \quad M_1 := P_1 \quad M_2 := \mathbb{1}\pi = \begin{bmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{bmatrix}$$

Proposition 5.1 (Baum formula for moments). *The binary hidden Markov map ϕ_n can be written in moment coordinates as*

$$\phi_n^\#(m_v) = \text{trace}(M_2 M_{v_1} M_{v_2} \cdots M_{v_n})$$

For example, $\phi_n^\#(m_{01001}) = \text{trace}(M_2 M_0 M_1 M_0 M_0 M_1)$.

Proof. By our definition of m_v (3), we have

$$\begin{aligned} \phi_n^\#(m_v) &= \sum_{w \geq v} \phi_n^\#(p_w) = \sum_{w \geq v} \text{trace}((\mathbb{1}\pi) P_{w_1} P_{w_2} \cdots P_{w_n}) \\ &= \text{trace} \left((\mathbb{1}\pi) \left(\sum_{w_1 \geq v_1} P_{w_1} \right) \left(\sum_{w_2 \geq v_2} P_{w_2} \right) \cdots \left(\sum_{w_n \geq v_n} P_{w_n} \right) \right) \\ &= \text{trace}(M_2 M_{v_1} M_{v_2} \cdots M_{v_n}) \end{aligned} \quad \square$$

5.3 Truncation and ψ_∞

Proposition 5.2. *The binary hidden Markov maps ϕ_n form a directed system of maps under marginalization, meaning that, for each $n' \geq n \geq 1$, the following diagrams commute:*

$$\begin{array}{ccc} & \mathbb{C}_m^{2^{n'}-1} & \\ \phi_{n'}^\# \nearrow & & \downarrow \mu_n^{n'} \\ \mathbb{C}_\theta^5 & & \mathbb{C}_m^{2^n-1} \\ \phi_n \searrow & & \end{array} \quad \begin{array}{ccc} & \bar{R}_{m,n'} & \\ \phi_{n'}^\# \nearrow & & \uparrow \mu_n^{n'\#} \\ \mathbb{C}[\theta] & & \bar{R}_{m,n} \\ \phi_n^\# \searrow & & \end{array}$$

Proof. This can be seen directly from the definition of ϕ_n using (1) and of m_v in (3). Alternatively, observe that because $M_0 = T$ is stochastic, $M_0 M_2 = M_2$, so for any sequence $\bar{0}$ of length $n' - n$, the Baum formula for moments (Proposition 5.1) implies that

$$\phi_{n'}^\#(m_{v\bar{0}}) = \phi_n^\#(m_v) \quad (10) \quad \square$$

³ P can be thought of naturally as a $2 \times 2 \times 2$ tensor, but we will not make use of this interpretation.

Thus, to compute ϕ_n for all n , it is only necessary to compute those $\phi_n^\# m_{v'}$ where v' ends in 1. Motivated by this observation, let $\overline{R}_{m,\infty} := \mathbb{C}[m_{v1} \mid v \in \{0,1\}^n \text{ for some } n \geq 0] = \mathbb{C}[m_1, m_{01}, m_{11}, m_{001}, m_{101}, m_{011}, \dots]$, which in subset index notation is simply

$$\begin{aligned} \overline{R}_{m,\infty} &:= \mathbb{C}[m_I \mid I \subseteq [n] \text{ for some } n \geq 0] \\ &= \mathbb{C}[m_1, m_2, m_{12}, m_3, m_{13}, m_{23}, \dots] \end{aligned}$$

Then we define $\phi_\infty : \mathbb{C}_\theta^5 \rightarrow \text{Spec } \overline{R}_{m,\infty}$ and $\phi_\infty^\# : \mathbb{C}[\theta] \leftarrow \overline{R}_{m,\infty}$ by the formula $\phi_\infty^\#(m_{v1\bar{0}}) := \phi_{\text{length}(v1)}^\#(m_{v1})$, i.e.

$$\phi_\infty^\#(m_I) := \phi_{\text{size}(I)}^\#(m_I) \quad (11)$$

Note that by locating the position of the last 1 in a binary sequence $v' \neq 0 \dots 0$, we can write v' in the form $v1\bar{0}$ for a unique string v (possibly empty if $v' = 1$), so this map is well-defined. By the same principle, for each n we can also define a “truncation” map $\tau : \text{Spec } \overline{R}_{m,\infty} \rightarrow \mathbb{C}_m^{2^n-1}$ by $\tau^\#(m_{v1\bar{0}}) := m_{v1}$, which, in subset index notation, is a literal ring inclusion:

$$\tau^\#(m_I) := m_I \quad (12)$$

With this definition, $\phi_n^\#$ factorizes as $\phi_n^\# = \phi_\infty^\# \circ \tau_n^\#$. Because we need to work with the new parameters $\eta_0 = (a_0, b, c_0, u, v_0)$, we define $\psi_\infty^\# := \chi^\# \circ \phi_\infty^\#$ so that $\psi_n^\# = \chi^\# \circ \phi_n^\# = \chi^\# \circ \phi_\infty^\# \circ \tau_n^\# = \psi_\infty^\# \circ \tau_n^\#$. We can summarize this and Proposition 5.2 in terms of ψ as follows:

Proposition 5.3. *For all $n' \geq n \geq 1$, the following diagrams commute:*

$$\begin{array}{ccc} & \mathbb{C}_{\eta_0}^5 & \\ \psi_n \swarrow & \downarrow \psi_{n'} & \searrow \psi_\infty \\ \mathbb{C}_m^{2^n-1} & \mathbb{C}_m^{2^{n'}-1} & \text{Spec } \overline{R}_{m,\infty} \\ \mu_n^{n'} \longleftarrow & & \longleftarrow \tau_{n'} \end{array} \quad \begin{array}{ccc} & \mathbb{C}[\eta_0] & \\ \psi_n^\# \nearrow & \uparrow \psi_{n'}^\# & \nwarrow \psi_\infty^\# \\ \overline{R}_{m,n} & \overline{R}_{m,n'} & \overline{R}_{m,\infty} \\ \mu_n^{n'} \longrightarrow & & \longrightarrow \tau_{n'}^\# \end{array}$$

Remark 5.4. These diagrams exhibit the rings $\overline{R}_{m,n}$ and maps $\psi_n^\#$ as a directed system under the inclusion maps $\mu_n^{n'}$, such that $\overline{R}_{m,\infty} = \text{colim}_{n \rightarrow \infty} \overline{R}_{m,n}$ and $\psi_\infty^\# = \lim_{n \rightarrow \infty} \psi_n^\#$.

Now, to prove that ψ_n factors through \mathfrak{q} , we need only show that ψ_∞ does.

5.4 Factoring ψ_∞ through a trace variety

Let X_0, X_1, X_2 be 2×2 matrices of indeterminates,

$$X_0 = \begin{bmatrix} x_{000} & x_{001} \\ x_{010} & x_{011} \end{bmatrix} \quad X_1 = \begin{bmatrix} x_{100} & x_{101} \\ x_{110} & x_{111} \end{bmatrix} \quad X_2 = \begin{bmatrix} x_{200} & x_{201} \\ x_{210} & x_{211} \end{bmatrix}$$

and following the notation of [Drensky, 2007], $\Omega_{2,3} := \mathbb{C}[\text{entries of } X_0, X_1, X_2]$ denotes the polynomial ring on the entries x_{ijk} of these three 2×2 matrices. The *trace algebra* $C_{2,3}$ is defined as the subring of $\Omega_{2,3}$ generated by the traces of products of these matrices, $C_{2,3} := \mathbb{C}[\text{trace}(X_{i_1} X_{i_2} \cdots X_{i_r}) \mid r \geq 1] \subseteq \Omega_{2,3}$ and we refer to $\text{Spec } C_{2,3}$ as a *trace variety*. We write

$$\nu : \text{Spec } \Omega_{2,3} \rightarrow \text{Spec } C_{2,3} \quad \text{and} \quad \nu^\# : C_{2,3} \hookrightarrow \Omega_{2,3}$$

for the natural dominant map and corresponding ring inclusion. To relate these varieties to binary HMMs, we define two new maps $\omega^\# : \Omega_{2,3} \rightarrow \mathbb{C}[\eta_0]$ and $\xi^\# : \overline{R}_{m,\infty} \rightarrow C_{2,3}$ by

$$\omega^\#(X_i) := \chi^\#(M_i) \quad \text{and} \quad \xi^\#(m_{v1}) := \text{trace} \left(\left(X_2 \prod_{i \in v} X_i \right) X_1 \right).$$

Proposition 5.5 (Baum factorization). *The ring map $\psi_\infty^\#$ factorizes as $\psi_\infty^\# = \omega^\# \circ \nu^\# \circ \xi^\#$, i.e., the following diagram commutes:*

$$\begin{array}{ccc} \mathbb{C}_{\eta_0}^5 & \xrightarrow{\psi_\infty} & \text{Spec } \overline{R}_{m,\infty} \\ \omega \downarrow & & \uparrow \xi \\ \text{Spec } \Omega_{2,3} & \xrightarrow{\nu} & \text{Spec } C_{2,3} \end{array}$$

Proof. This is just a restatement of the Baum formula for moments (Proposition 5.1):

$$\begin{aligned} \omega^\#(\nu^\#(\xi^\#(m_{v1}))) &= \omega^\# \text{trace} \left(X_2 \prod_{i \in v1} X_i \right) = \chi^\# \text{trace} \left(M_2 \prod_{i \in v1} M_i \right) \\ &= \chi^\#(\phi^\#(m_{v1})) = \psi_\infty^\#(m_{v1}) \end{aligned} \quad \square$$

5.5 Proving the Parameter Subring Lemma (4.2)

We begin by seeking a factorization of the map $\omega^\# \circ \nu^\#$. For this we apply the following commutative algebra result of Sibirskii on the trace algebras $C_{2,r}$:

Proposition 5.6 (Sibirskii, 1968). *The trace algebra $C_{2,r}$ is generated by the elements*

$$\begin{aligned} \text{trace}(X_i) &: 0 \leq i \leq r \\ \text{trace}(X_i X_j) &: 0 \leq i \leq j \leq r \\ \text{trace}(X_i X_j X_k) &: 0 \leq i < j < k \leq r \end{aligned}$$

Corollary 5.7. *The algebra $C_{2,3}$ is generated by the 10 elements*

$$\begin{aligned} &\text{trace}(X_0), \text{trace}(X_1), \text{trace}(X_2), \\ &\text{trace}(X_0^2), \text{trace}(X_1^2), \text{trace}(X_2^2), \text{trace}(X_0 X_1), \text{trace}(X_0 X_2), \text{trace}(X_1 X_2), \\ &\text{trace}(X_0 X_1 X_2) \end{aligned}$$

Proposition 5.8. *The ring map $\omega^\# \circ \nu^\#$ factors through the inclusion*

$$\mathfrak{q}^\# : \mathbb{C}[\eta] := \mathbb{C}[a, b, c, u, v] \hookrightarrow \mathbb{C}[\eta_0] := \mathbb{C}[a_0, b, c_0, u, v_0],$$

i.e. we can write $\omega^\# \circ \nu^\# = \mathfrak{q}^\# \circ \mathfrak{r}^\#$ so that the following diagram commutes:

$$\begin{array}{ccc} \mathbb{C}_{\eta_0}^5 & \xrightarrow{\mathfrak{q}} & \mathbb{C}_\eta^5 \\ \omega \downarrow & & \downarrow \mathfrak{r} \\ \text{Spec } \Omega_{2,3} & \xrightarrow{\nu} & \text{Spec } C_{2,3} \end{array}$$

Proof. We apply $\omega^\#$ to the ten generators of $C_{2,3}$ given in Corollary 5.7 and check that they land in $\mathbb{C}[\eta]$. Explicit, we find that:

$$\begin{aligned} \text{trace}(\chi^\# M_0) &= b + 1 & \text{trace}(\chi^\# M_1) &= bu + c + u & \text{trace}(\chi^\# M_2) &= 1 \\ \text{trace}(\chi^\# M_0^2) &= b^2 + 1 & \text{trace}(\chi^\# M_1^2) &= b^2 u^2 + 2bcu + c^2 + 2cu + u^2 + 2bv \\ \text{trace}(\chi^\# M_2^2) &= 1 & \text{trace}(\chi^\# M_0 M_1) &= b^2 u + bc + c + u & \text{trace}(\chi^\# M_0 M_2) &= 1 \\ \text{trace}(\chi^\# M_1 M_2) &= a + u & \text{trace}(\chi^\# M_0 M_1 M_2) &= ab + c + u & & \square \end{aligned}$$

Now, by letting $\overline{\psi}_\infty^\# := \mathfrak{r}^\# \circ \xi^\#$ we may factor the ring map $\psi_\infty^\#$ as

$$\psi_\infty^\# = \omega^\# \circ \nu^\# \circ \xi^\# = \mathfrak{q}^\# \circ \mathfrak{r}^\# \circ \xi^\# = \mathfrak{q}^\# \circ \overline{\psi}_\infty^\#.$$

Corollary 5.9. *The following diagram commutes:*

$$\begin{array}{ccccc} & & \psi_\infty & & \\ & & \curvearrowright & & \\ \mathbb{C}_{\eta_0}^5 & \xrightarrow{\mathfrak{q}} & \mathbb{C}_\eta^5 & \xrightarrow{\overline{\psi}_\infty} & \text{Spec } \overline{R}_{m,\infty} \\ \omega \downarrow & & \downarrow \mathfrak{r} & \nearrow \xi & \\ \text{Spec } \Omega_{2,3} & \xrightarrow{\nu} & \text{Spec } C_{2,3} & & \end{array}$$

Proof of the Parameter Subring Lemma (4.2). Proposition 5.3 and Corollary 5.9 together imply that the following diagrams commute:

$$\begin{array}{ccccccc} \mathbb{C}_{\eta_0}^5 & \xrightarrow{\mathfrak{q}} & \mathbb{C}_\eta^5 & \xrightarrow{\overline{\psi}_\infty} & \text{Spec } \overline{R}_{m,\infty} & \xrightarrow{\tau_n} & \mathbb{C}_m^{2^n-1} \\ & & & & & \searrow & \\ & & & & & \psi_n & \\ & & & & & \curvearrowright & \\ \mathbb{C}[\eta_0] & \xleftarrow{\mathfrak{q}^\#} & \mathbb{C}[\eta] & \xleftarrow{\overline{\psi}_\infty^\#} & \overline{R}_{m,\infty} & \xleftarrow{\tau_n^\#} & \overline{R}_{m,n} \\ & & & & & \searrow & \\ & & & & & \psi_n^\# & \end{array}$$

In particular, the map $\psi_n^\#$ factors through $\mathbb{C}[\eta]$, as required. \square

5.6 Proving the Birational Parameter Theorem (4.1)

Recall that Corollary 4.3 implies the first part of the Birational Parameter Theorem (4.1), by taking

$$\mathfrak{q} \circ \chi^{-1} : \mathbb{C}_\theta^5 \longrightarrow \mathbb{C}_\eta^5$$

as the generically 2 : 1 map. Thus, it remains to show that the maps

$$\overline{\psi}_n : \mathbb{C}_\eta^5 \longrightarrow \mathbf{V}_{\text{BHM}(n)}$$

have birational inverses ρ_n . The inverse map ρ_3 was already exhibited in Corollary 4.8, and we obtain ρ_n by marginalization: let

$$\rho_n = \rho_3 \circ \mu_3^n.$$

Let $U \subseteq \mathbb{C}_\eta^5$ be the Zariski open set on which $\bar{\psi}_3$ is an isomorphism with inverse ρ_3 . Consider the set $\bar{\psi}_n(U) \subseteq \mathbf{V}_{\mathbf{BHM}(n)}$. It is Zariski dense in $\mathbf{V}_{\mathbf{BHM}(n)}$, and by Chevalley's theorem (Grothendieck and Dieudonné, 1966, EGA IV, 1.8.4), it is constructible, so it must contain a dense open set $W' \subseteq \mathbf{V}_{\mathbf{BHM}(n)}$. Now let $W = \bar{\psi}_n^{-1}(W')$, so we have $\bar{\psi}_n(W) = W' \subseteq \bar{\psi}_n(U)$.

Proposition 5.10. $\rho_n \circ \bar{\psi}_n = \text{Id}$ on W and $\bar{\psi}_n \circ \rho_n = \text{Id}$ on W' .

Proof. Suppose $\hat{\eta} \in W$. Then $\rho_n \circ \bar{\psi}_n(\hat{\eta}) = \rho_3 \circ \mu_3^n \circ \bar{\psi}_n(\hat{\eta}) = \rho_3 \circ \bar{\psi}_3(\hat{\eta}) = \hat{\eta}$ since $\hat{\eta} \in U$. Now suppose $\hat{p} \in W'$, so $\hat{p} = \bar{\psi}_n(\hat{\eta})$ for some $\hat{\eta} \in W$. Then, applying Proposition 5.2,

$$\begin{aligned} \bar{\psi}_n \circ \rho_n(\hat{p}) &= \bar{\psi}_n \circ \rho_n \circ \bar{\psi}_n(\hat{\eta}) = \bar{\psi}_n \circ \rho_3 \circ \mu_3^n \circ \bar{\psi}_n(\hat{\eta}) \\ &= \bar{\psi}_n \circ \rho_3 \circ \bar{\psi}_3(\hat{\eta}) = \bar{\psi}_n(\hat{\eta}) = \hat{p} \end{aligned} \quad \square$$

This completes the proof of the Birational Parameter Theorem (4.1). In fact we have also proven the following:

Theorem 5.11. For each $n \geq 3$, there is a commutative diagram of dominant maps:

$$\begin{array}{ccccc} & & & \psi_{n+1} & \\ & & & \curvearrowright & \\ & & & & \mathbf{V}_{\mathbf{BHM}(n+1)} \\ & & & \bar{\psi}_{n+1} & \uparrow \\ \mathbb{C}_\theta^5 & \xrightarrow{\chi^{-1}} & \mathbb{C}_{\eta_0}^5 & \xrightarrow{\mathfrak{q}} & \mathbb{C}_\eta^5 & \xrightarrow{\bar{\psi}_{n+1}} & \mathbf{V}_{\mathbf{BHM}(n+1)} \\ & & & \bar{\psi}_n & \searrow & \downarrow \mu_n^{n+1} \\ & & & \psi_n & \curvearrowleft & \mathbf{V}_{\mathbf{BHM}(n)} \end{array}$$

6 Applications and future directions

Besides attempting to compute a set of generators for $\mathbf{I}_{\mathbf{BHM}(5)}$, there are many other questions to be answered about HMMs that can be approached immediately with the techniques of this paper.

6.1 A nonnegative distribution in $\mathbf{V}_{\mathbf{BHM}(3)}$ but not $\mathbf{M}_{\mathbf{BHM}(3)}$

It turns out that not all the non-negative real points of $\mathbf{V}_{\mathbf{BHM}(n)}$ lie in the model $\mathbf{M}_{\mathbf{BHM}(n)}$. The following real point $\hat{\theta}$ of \mathbb{C}_θ^5 does not lie in Δ_θ^5 , but maps under ϕ_3 to a point \hat{p} of Δ_p^7 :

$$\hat{\theta} = (\hat{\pi}, \hat{T}, \hat{E}) = \left(\left[-\frac{1}{8} \quad \frac{9}{8} \right], \quad \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}, \quad \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \right) \quad (13)$$

Moreover, the analysis of Section 4.3 reveals that the fiber $\phi_3^{-1}(\hat{p})$ consists only of the point $\hat{\theta}$ and the “swapped” point

$$\hat{\theta}' = (\hat{\pi}', \hat{T}', \hat{E}') = \left(\left[\frac{9}{8} \quad -\frac{1}{8} \right], \quad \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}, \quad \begin{bmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix} \right) \quad (14)$$

which is also not in Δ_{θ}^5 . Hence the image point $\widehat{p} = \phi_3(\widehat{\theta}) = \phi_3(\widehat{\theta}')$ is a non-negative point of $\mathbf{V}_{\mathbf{BHM}(3)}$ that does not lie in $\mathbf{M}_{\mathbf{BHM}(3)}$.

6.2 Testing model membership

In light of the fact that not every nonnegative distribution in $\mathbf{V}_{\mathbf{BHM}(n)}$ is in $\mathbf{M}_{\mathbf{BHM}(n)}$, the defining equations of $\mathbf{V}_{\mathbf{BHM}(n)}$ are not sufficient to test a probability distribution for membership to the model.

So, suppose we are given a distribution $p \in \Delta_p^{2^n-1}$ and asked to determine if $p \in \mathbf{M}_{\mathbf{BHM}(n)}$. The following procedure yields either

- (1) a proof by contradiction that $p \notin \mathbf{M}_{\mathbf{BHM}(n)}$,
- (2) a parameter vector $\theta \in \Delta_{\theta}^5$ such that $\phi_n(\theta) = p \in \mathbf{M}_{\mathbf{BHM}(n)}$, or
- (3) a reduction of the question to whether p lies in one of the lower-dimensional submodels of $\mathbf{M}_{\mathbf{BHM}(n)}$ discussed in Section 4.3.

How to proceed from (3) is not difficult, and will be discussed explicitly in future work.

To begin, we let $p' = \mu_3^n(p) \in \Delta_p^{2^3-1}$, i.e. we marginalize p to the distribution p' it induces on the first three visible nodes. Note that if $p \in \mathbf{M}_{\mathbf{BHM}(n)}$ then $p' \in \mathbf{M}_{\mathbf{BHM}(3)}$. Observing the moments m_I of p' , if any denominators in the formulae of Corollary 4.8 vanish, then we end in case (3).

Otherwise, we let $(a, b, c, u, v) = \overline{\psi}_3^{-1}(p')$, choose v_0 to be either square root of v , and let $a_0 = a/v_0$, $c_0 = c/v_0$. If p were due to some BHM process, then by Theorem 5.11, these would be its parameters, up to a simultaneous sign change of (a_0, b_0, v_0) . With this in mind, we define $\theta = (\pi, T, E)$ using (9). If (π, T, E) are not non-negative stochastic matrices, then $p \notin \mathbf{M}_{\mathbf{BHM}(n)}$ and we end in case (1). If they are, we compute $p'' = \phi_n(\theta)$, and if $p = p''$ then we end in case (2). Otherwise p must not have been in $\mathbf{M}_{\mathbf{BHM}(n)}$, so we end in case (1).

Note that since all the criteria in this test are algebraic equalities and inequalities, this procedure implicitly describes a semialgebraic characterization of $\mathbf{M}_{\mathbf{BHM}(n)}$ for all $n \geq 3$.

6.3 Identifiability of parameters

By a *rational map* on a possibly non-algebraic subset $\Theta \subseteq \mathbb{C}^k$, we mean any rational map on the Zariski closure of Θ , which will necessarily be defined as a function on a Zariski dense open subset of Θ . We define polynomial maps on Θ similarly.

Let $\phi : \Theta \rightarrow \mathbb{C}^n$ be an algebraic statistical model, where as usual we assume $\Theta \subseteq \mathbb{C}^k$ is Zariski dense, and therefore Zariski irreducible. A (rational) parameter of the model is any rational map $s : \Theta \rightarrow \mathbb{C}$. Such parameters form a field, $K \simeq \text{Frac}(\mathbb{C}^k)$. In applications such as [Meshkat, Eisenberg, and DiStefano, 2009], it is important to know to what extent a parameter can be identified from observational data alone. In other words, given $\phi(\theta)$, what can we say about $s(\theta)$? This leads to several different notions of parameter identifiability, as discussed by Sullivant, Garcia-Puente, and Spielvogel [2010].

Definition 6.1. We say that a rational parameter $s \in K$ is

- (*set-theoretically identifiable*) if $s = \sigma \circ \phi$ for some set-theoretic function $\sigma : \phi(\Theta) \rightarrow \mathbb{C}$. In other words, for all $\theta, \theta' \in \Theta$, if $\phi(\theta) = \phi(\theta')$ then $s(\theta) = s(\theta')$.
- (*rationally identifiable*) if $s = \sigma \circ \phi$ for some *rational* map $\sigma : \phi(\Theta) \rightarrow \mathbb{C}$ (this notion is used without a name by Sullivant et al. [2010]).
- (*generically identifiable*) if there is a (relatively) Zariski dense open subset $U \subseteq \Theta$ such that $s|_U = \sigma \circ \phi|_U$ for some set-theoretic function $\sigma : \phi(U) \rightarrow \mathbb{C}$.

- *algebraically identifiable* if there is a polynomial function $g(p, q) := \sum_i g_i(p_1, \dots, p_n)q^i$ on $\phi(\Theta) \times \mathbb{C}$ of degree $d > 0$ in q (so that g_d is not identically 0 on $\phi(\Theta)$) such that $g(\phi(\theta), s(\theta)) = 0$ for all $\theta \in \Theta$ (and hence all $\theta \in \mathbb{C}^k$).

Question 6.2. What combinations of BHM parameters are rationally identifiable, generically identifiable, or algebraically identifiable?

To answer this question we introduce a lemma on algebraic statistical models in general:

Lemma 6.3. *For any algebraic statistical model ϕ as above, the sets K_{ri} , K_{gi} , and K_{ai} , of rationally, generically, and algebraically identifiable parameters, respectively, are all fields.*

Proof. Since Θ is Zariski irreducible, so is $\phi(\Theta)$. Hence the set of rational maps on $\phi(\Theta)$ is simply the fraction field of its Zariski closure (an irreducible variety), and K_{ri} is the image of this field under $\phi^\#$, which must be a field.

For K_{gi} , the crux is to show that if $s, s' \in K_{gi}$ and $s \neq 0$ then $s'/s \in K_{gi}$. Let $U \subseteq \Theta$ and $\sigma : \phi(U) \rightarrow \mathbb{C}$ be as in the definition for s , and likewise $U' \subseteq \Theta$ and $\sigma' : \phi(U') \rightarrow \mathbb{C}$ for s' . Let $U'' = \{\theta \in U \cap U' \mid s(\theta) \neq 0\}$, which, being an intersection of three Zariski dense open subsets of Θ , is a dense open. We have $\sigma' \neq 0$ on $\phi(U'') \subseteq \phi(U) \cap \phi(U')$, so we can let $\sigma'' = \sigma'/\sigma : \phi(U'') \rightarrow \mathbb{C}$, and then $\sigma'' \circ \phi = s'/s$, so $s'/s \in K_{gi}$. Thus K_{gi} is stable under division, and simpler arguments show it is stable under $+$, $-$, and \cdot , so it is a field.

Finally, K_{ai} is expressly the relative algebraic closure in K of the image under $\phi^\#$ of the coordinate ring of $\phi(\Theta)$, which is therefore a field. \square

Proposition 6.4. *For any algebraic statistical model ϕ as above, $K_{ri} \subseteq K_{gi} \subseteq K_{ai} \subseteq K$.*

Proof. This is now just a restatement of Proposition 3 in [Sullivant et al., 2010]. \square

Now, the answer to our identifiability question for BHM parameters can be given easily in the coordinates of Section 4. Here ϕ is the BHM map ϕ_n . The field K_{ri} is simply the image $\phi^\#(\text{Frac}(\mathbb{C}_\eta^5))$ because by Theorem 4.1,

$$\overline{\psi}^\# : \text{Frac}(\mathbf{V}_{\text{BHM}(n)}) \rightarrow \text{Frac}(\mathbb{C}_\eta^5)$$

is an isomorphism. Hence the rationally identifiable parameters are precisely the field of rational functions in $(a, b, c, u, v) = (a_0 v_0, b, c_0 v_0, u, v_0^2)$ (see (9) for the meanings of these parameters). Since K is a quadratic field extension of K_{ri} given by adjoining $v_0 = \sqrt{v}$, and K_{ai} is the algebraic closure of K_{ri} in K (almost by definition), it follows that $K_{ai} = K$, i.e. *all parameters* are algebraically identifiable. Finally, we observe that, by the action of **sw** in Section 4.1, there are generically two possible values of $v_0 = \frac{1}{2}(E_{11} - E_{01})$ for a given observed distribution, namely $\pm\sqrt{v}$. Hence $v_0 \notin K_{gi}$, and since a quadratic field extension has no intermediate extensions, it follows that $K_{ri} = K_{gi}$, i.e. all generically identifiable parameters are in fact rationally identifiable. In summary,

Proposition 6.5. *For $\mathbf{M}_{\text{BHM}(n)}$ where $n \geq 3$,*

$$\mathbb{C}(a, b, c, u, v) = K_{ri} = K_{gi} \subsetneq K_{ai} = \mathbb{C}(a_0, b, c_0, u, v_0)$$

6.4 A new grading on BHMM invariants

The re-parametrized model map $\overline{\psi}_n$ is homogeneous in cumulant and moment coordinates, with respect to a \mathbb{Z} -grading where $\text{deg}(m_v) = \text{deg}(k_v) = \text{sum}(v)$, $\text{deg}(b) = 0$, $\text{deg}(a) = \text{deg}(c) = \text{deg}(u) = 1$, and $\text{deg}(v) = 2$. This grading allows for fast linear algebra techniques

that solve for low degree model invariants as in [Bray and Morton, 2005], except that this grading is intrinsic to the model. Bray and Morton’s grading, which is in *probability* coordinates, is not on the binary HMM proper, but on a larger variety obtained by relaxing the parameter constraints that the transition and emission matrix row sums are 1. The invariants obtained in their search are hence invariants of this larger variety, and exclude some invariants of $\mathbf{M}_{\text{BHM}(n)}$. The grading presented here can thus be used to complete their search for invariants up to any finite degree.

6.5 Equilibrium BHM processes

In Section 4.3 we found that if a BHM process is at equilibrium, our formula for $\overline{\psi}_3^{-1}$ is undefined. We may define Equilibrium Binary Hidden Markov Models, EBHMMs, by restricting ψ_n to the locus $\{a_0b - a_0 + c_0 = 0\}$, which turns out to yield a four-dimensional submodel of $\mathbf{M}_{\text{BHM}(n)}$ for each $n \geq 3$. The same techniques used here to study BHMMs reveal that the EBHMMs, too, have birational parametrizations, and when $n = 3$, the EBHMM has even simpler defining equations than the BHMM. The geometry of EBHMMs will need to be considered explicitly in future work to identify the learning coefficients of BHMM fibers.

6.6 HMMs with more visible states

The results of this paper can be used directly to identify the parameters of a generic hidden Markov process with two hidden states and k visible states $\alpha_1, \dots, \alpha_k$. Such a process can be specified by a $2 \times k$ matrix E of emission probabilities, along with a triple (a_0, b, c_0) defining the π and T of the two-state hidden Markov chain as in (9).

To obtain E_{0j} and E_{1j} from the observed probability distribution for any fixed j , we simply define a BHM process by letting $\alpha_j = 1$ and $\alpha_i = 0$ for $i \neq j$. Applying Proposition 4.7 to the moments of the distribution yields values for (a, b, c, u, v) provided the genericity condition that the denominators involved do not vanish. Letting $v_0 = \sqrt{v}$, $a_0 = a/v_0$, and $c_0 = c/v_0$, we obtain (a_0, b, c_0, u, v_0) up to a simultaneous sign change on (a_0, c_0, v_0) corresponding to swapping the hidden alphabet as in Section 4.1. Then $E_{0j} = u - v$ and $E_{1j} = u + v$, and we get π, T as well from (a_0, b, c_0) . We can repeat this for each $j = 1, \dots, k$ to obtain all the emission parameters, and hence identify all the process parameters modulo the swapping operation. With any luck, perhaps similar techniques could elucidate the algebraic statistics and geometry of HMMs with any number of hidden states.

References

- J. Baker. The DRAGON system – An overview. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(1):24 – 29, Feb 1975.
- D. J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler. Bertini: Software for numerical algebraic geometry. Available at <http://www.nd.edu/~sommese/bertini>.
- D. J. Bates, J. D. Hauenstein, C. Peterson, and A. J. Sommese. *Numerical decomposition of the rank-deficiency set of a matrix of multivariate polynomials*, pages 55–77. Texts and Monographs in Symbolic Computation. Springer-Verlag, 2010.
- L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37:1554–1563, 1966.

- N. Bray and J. Morton. Equations defining hidden Markov models. In *Algebraic Statistics for Computational Biology*, chapter 11. Cambridge University Press, 2005.
- M. Casanellas and J. Fernandez-Sanchez. Performance of a new invariants method on homogeneous and non-homogeneous quartet trees. *Molecular Biology and Evolution*, 24:288–293, 2006.
- D. A. Cox, J. B. Little, and D. O’Shea. *Ideals, Varieties, and Algorithms, Third Edition*. Springer New York, 2007.
- P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Stat.*, 26(1):363–397, 1998.
- V. Drensky. Computing with matrix invariants. *Math. Balkanica*, 21:141–172, 2007.
- M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*. Oberwolfach Seminars 39, 2009.
- N. Eriksson. Using invariants for phylogenetic tree construction. In *Emerging Applications of Algebraic Geometry*. I.M.A. Volumes in Mathematics and its Applications, 2008.
- D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at <http://www.math.uiuc.edu/Macaulay2/>.
- A. Grothendieck and J. Dieudonné. *Éléments de géométrie algébrique (rédigés avec la collaboration de Jean Dieudonné) IV. Étude locale des schémas et des morphismes de schémas, Troisième partie*. Publications Mathématiques de l’IHÉS, 1966.
- A. Krogh, I. S. Mian, and D. Haussler. A Hidden Markov Model that finds genes in E. coli DNA. *Nucleic Acids Research*, page 47684778, 1994.
- N. Meshkat, M. Eisenberg, and J. J. DiStefano. An algorithm for finding globally identifiable parameter combinations of nonlinear ode models using Gröbner bases. *Mathematical Biosciences*, 222(2):61 – 72, 2009.
- L. Pachter and B. Sturmfels. *Algebraic Statistics for Computational Biology*. Cambridge University Press, 2005.
- G. Pistone, E. Riccomagno, and H. P. Wynn. *Computational commutative algebra in discrete statistics*. Chapman and Hall / CRC, 2001.
- A. Schönhuth. Generic identification of binary-valued hidden Markov processes, 2011. arXiv:1101.3712.
- K. Sibirskii. Algebraic invariants for a set of matrices. *Siberian Mathematical Journal*, 9: 115–124, 1968. ISSN 0037-4466.
- J. Q. Smith and P. Zwiernik. Tree cumulants and the geometry of binary tree models, 2010. arXiv:1004.4360v3.
- R. L. Stratonovich. Conditional Markov Processes. *Theory of Probability and its Applications*, pages 156–178, 1960.
- B. Sturmfels and P. Zwiernik. Binary cumulant varieties, 2011. arXiv:1103.0153.

- S. Sullivant, L. D. Garcia-Puente, and S. Spielvogel. Identifying causal effects with computer algebra. Proceedings of the 26th Conference of Uncertainty in Artificial Intelligence, 2010.
- S. Watanabe. *Algebraic Geometry and Statistical Learning Theory (Cambridge Monographs on Applied and Computational Mathematics)*. Cambridge University Press, 2009.
- B.-J. Yoon. Hidden markov models and their applications in biological sequence analysis. *Current Genomics*, 10(6):402–415, 2009.