Identifiability of Product of Experts Models

Spencer L. Gordon
U. of Liverpool

Manav Kant
Caltech

Eric Y. Ma
Caltech

Leonard J. Schulman
Caltech

Andrei C. Staicu
Caltech

Abstract

Product of experts (PoE) are layered networks in which the value at each node is an AND (or product) of the values (possibly negated) at its inputs. These were introduced as a neural network architecture that can efficiently learn to generate high-dimensional data which satisfy many low-dimensional constraints—thereby allowing each individual expert to perform a simple task. PoEs have found a variety of applications in learning. We study the problem of identifiability of a product of experts model having a layer of binary latent variables, and a layer of binary observables that are iid conditional on the latents. The previous best upper bound on the number of observables needed to identify the model was exponential in the number of parameters. We show: (a) When the latents are uniformly distributed, the model is identifiable with a number of observables equal to the number of parameters (and hence best possible). (b) In the more general case of arbitrarily distributed latents, the model is identifiable for a number of observables that is still linear in the number of parameters (and within a factor of two of best-possible). The proofs rely on root interlacing phenomena for some special three-term recurrences.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

Product of Experts Models In modeling complex, high-dimensional data, it is often necessary to combine various simple distributions to produce a more expressive distribution. One way of doing this is the mixture model, or weighted sum of distributions. Alone, however, this still requires quite expressive components, which is a hindrance for modeling data in a high-dimensional space. Product of experts (PoE) were introduced in the neural networks literature as an antidote to this problem: the distribution is factorized into a set of independent lower-dimensional distributions [Hinton, 1999, Hinton, 2002]. Equivalently, the overall distribution is an AND over the factor distributions (which may themselves be mixture models).

PoEs have recently been applied to solve diverse problems requiring the generation of data that simultaneously satisfy numerous sets of constraints. For example, the PoE-GAN framework has advanced the state of the art in multimodal conditional image synthesis, generating images conditioned on all or some subset of text, sketch, and segmentation inputs [Huang et al., 2022]. In the field of language generation, a PoE with two factors, a pre-trained language model and a combination of a toxicity expert with an anti-expert, was used to steer a language model away from offensive outputs [Liu et al., 2021].

However, fundamental questions about PoE models remain unresolved. In this paper, we study a PoE in which each observable random variable (rv) X depends upon statistically independent latent rv's $\mathcal{U} = (U_1, \ldots, U_\ell)$. (We use a slightly nonstandard "latent variable" parameterization of the PoE; see Sec. 4 for the correspondence between this framework and the standard one.)

The *model* here is the prior distributions on the U_j and the likelihoods $\Pr(X = x|\mathcal{U})$. We investigate the

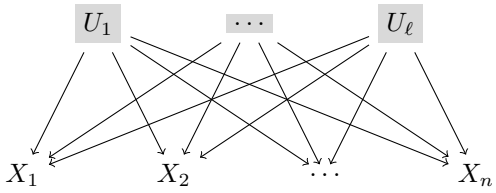

Figure 1: The graphical model (latent variables shaded)

well-studied class of instances in which the X_i and the U_j are binary and $\Pr(X = 1|\mathcal{U})$ can be expressed as a product over the U_j's, namely, for some coefficients α,

$$\Pr(X = 1 \mid \mathcal{U} = u) = \prod_{j=1}^{\ell} \alpha_{j,u_j} \tag{1}$$

This class of models has a symmetry group $\mathcal{G}$ involving its latent parameters: the distribution of X conditional on $\mathcal{U}$ is invariant to permutation of the $\mathcal{U}_j$'s, and also to continuous gauge transformations of the parameters (which we spell out in Sections 2.1, 3.1); due to the latter, $\mathcal{G}$ has positive dimension as a real manifold.

We are concerned with the following question: what is the minimum n needed so that instances of this class can be identified from the statistics of n independent samples $X_1, \ldots, X_n$? (Specifically we focus on local identification, see below.) It is necessary to sample at least as many observable variables as there are parameters, minus the dimension of $\mathcal{G}$ as a manifold. However, the previous best upper bound on n was exponential in the number of parameters.

We show that (a) In the case of uniform priors on the U_j's, local identifiability holds with the least possible number of observables. (b) In the case of general priors, local identifiability holds with at most twice the least possible number of observables. This resolves the previous exponential gap that existed for both versions of this problem.

Algebraic Mappings The model described above corresponds to a directed graphical model, or Bayesian network, with all edges directed from latent toward observable vertices (see Fig. 1). The independence of the U_j's in the prior is implicit in that these vertices are sources in this graph. The representation (1) specifies an algebraic mapping from model parameters (namely, the α's and the prior distributions on the U_j) to a probability distribution on $\mathcal{X}$.

In Section 2 the prior on each U_j is assumed uniform, so the mapping we are concerned with is from the α's to the distribution of $\mathcal{X}$. Section 3 treats the general case of arbitrary priors on each U_j.

Identifiability of a map f. If the preimage of every point in $\mathrm{im}(f)$, apart from a set of measure 0 called critical values, is a finite union of orbits under $\mathcal{G}$, we say the model is locally identifiable. If, further, at non-critical values the pre-image is a single orbit under $\mathcal{G}$, then we say that the model is fully identifiable. In this paper, we consider a family of polynomials $f : \mathbb{R}^m \to \mathbb{R}^m$ that (after first quotienting out the continuous degrees of freedom of $\mathcal{G}$) map parameters to observable statistics. By Sard's theorem, the set of critical values is of measure 0 in the codomain; therefore, if f is anywhere a local isomorphism, then the preimage of any point, except for the measure-0 set of critical values, is a union of isolated points. The preimage of any point is, however, a real variety; it must therefore be a 0-dimensional real variety, and is therefore (by the Hilbert basis theorem) a finite point set.

Thus, in order to show local identifiability, it will suffice in the remainder of the paper to show that there is a point in the domain at which the Jacobian of f is non-singular.

The question of local identifiability along with the related question of whether the model generates all possible distributions on $\mathcal{X}$, are the subjects of much of the extensive literature about this model and its undirected-graph variant (see below) in the neural networks and algebraic statistics literatures [Hinton et al., 2006, Drton et al., 2007, Roux and Bengio, 2008, Cueto et al., 2010, Long and Servedio, 2010, Hinton, 2010, Martens et al., 2013, Montúfar and Morton, 2015, Montúfar and Morton, 2017, Montúfar, 2018, Seigal and Montúfar, 2018]. We encountered this question in the context of causal identification and the classical moment problem [Gordon et al., 2020, Gordon et al., 2021, Gordon et al., 2023].

The only prior upper bound for the number of observables required for model identification, was achieved by treating $\mathcal{U}$ as a single "lumped" rv, taking on 2^ℓ possible values. Then there is a longstanding result [Blischke, 1964] implying an upper bound on n of $2^{\ell+1} - 1$. This is exponentially larger than the number of parameters of the model.

We close this exponential gap, both for uniform and for general priors. We introduce a general method for showing local identifiability of these latent symmetric models. Our main results are:

Theorem 1: In the uniform-prior case, the model can be locally identified with $n = \ell + 1$ observables, which is best possible, as it matches the number of degrees of freedom of the model (after quotienting-out

symmetries).

Theorem 16: In the general-prior case, the model can be locally identified with $n \leq 4\ell + 1$ observables, which is within a factor of two of best possible.

These theorems are logically incomparable. Their proofs share a basic structure, but the second involves some surprising ingredients that are not foreshadowed in the first.

Since the running time of any algorithm for identification of PoEs from this class, is lower bounded by the number of moments required for identification, our results open up the possibility of a fast algorithm for the identification problem.

2 IDENTIFYING MODELS WITH UNIFORM PRIORS

2.1 Preliminaries

The Model In this section we treat the case that the prior on each U_j is uniform. Then the model has 2ℓ parameters α_{jb} for $j = 1, \ldots, \ell$ and $b = 0, 1$; and (1) yields that

$$\Pr(X = 1) = \boldsymbol{\mu}_1 \qquad \text{where}$$
$$\boldsymbol{\mu}_1 \coloneqq \prod_{j=1}^{\ell} \frac{\alpha_{j0} + \alpha_{j1}}{2} \tag{2}$$

and more generally

$$\Pr(X_1 = \ldots = X_t = 1) = \boldsymbol{\mu}_t \qquad \text{where}$$
$$\boldsymbol{\mu}_t \coloneqq \prod_{j=1}^{\ell} \frac{\alpha_{j0}^t + \alpha_{j1}^t}{2}. \tag{3}$$

Hadamard Products of Hankel Matrices It was observed in [Cueto et al., 2010] that the RBM model under consideration there, was a Hadamard product of simpler models, each having only a single latent variable. A similar phenomenon occurs in the directed model we study.

For each $j \in [\ell]$, define $\mathrm{Han}(n, j)$ to be the $(n+1) \times (n+1)$ matrix with entries

$$\mathrm{Han}(n, j)_{ab} = (\alpha_{j0}^{a+b} + \alpha_{j1}^{a+b})/2 \qquad (0 \leq a, b \leq n).$$

This is a Hankel matrix of rank 2. The Hadamard (or entrywise) product of these matrices, ranging over $1 \leq j \leq \ell$, is

$$H(n)_{ab} = \boldsymbol{\mu}_{a+b} \qquad (0 \leq a, b \leq n).$$

If we had access to each individual Hankel matrix $\mathrm{Han}(2, j)$, we could apply the classical method of Prony or similar methods [de Prony, 1795, Gordon et al., 2020, Kim et al., 2019] to determine α_{j0} and α_{j1}; this method works (for an arbitrary prior on U_j) provided the Hankel matrix is rank-deficient.

However, what we have access to is only $H(n)$ and not the individual $\mathrm{Han}(n, j)$'s. One may consider $H(n)$ itself as a Hankel matrix, by ignoring the product structure on the U_j's and regarding U as a single latent variable with range $[2^\ell]$; but then in order to have rank deficiency, one requires the very large Hadamard matrix $H(2^\ell)$, and consequently, the Prony method is only applicable to identifying the model if we obtain $n = 2^\ell$ observables X_i. This is exponentially larger than the number of degrees of freedom of the model, which as we see from the model (3) is merely linear in ℓ. As noted earlier, this exponential gap was the motivation for our investigation. We show that the linear answer is correct: the model is locally identifiable as soon as n matches the number of degrees of freedom of the model (after quotienting out a continuous symmetry which we now describe).

Symmetries of the Model The model has both discrete and continuous symmetries. The moments $\boldsymbol{\mu}_t$ are invariant to:

1. *Discrete symmetries.* The wreath product $S_2 \wr S_\ell$ (a.k.a. hyperoctahedral group):
   (a) For any j, exchange α_{j0} and α_{j1}.
   (b) Exchange any j and j'. That is, for $j \neq j'$, exchange α_{j0} with $\alpha_{j'0}$, and α_{j1} with $\alpha_{j'1}$.
2. *Continuous symmetries.* For any $j \neq j'$ and $\lambda > 0$, the gauge transformation

$$\begin{cases} (\alpha_{j0}, \alpha_{j1}) & \mapsto (\lambda\alpha_{j0}, \lambda\alpha_{j1}), \\ (\alpha_{j'0}, \alpha_{j'1}) & \mapsto (\lambda^{-1}\alpha_{j'0}, \lambda^{-1}\alpha_{j'1}) \end{cases} \tag{4}$$

The model can of course be identified only up to these symmetries. We can therefore, w.l.o.g., take advantage of the gauge symmetries to scale the parameters α_{jb} so that for all $j \in [\ell]$,

$$\alpha_{j0} + \alpha_{j1} = 2\boldsymbol{\mu}_1^{1/\ell} =: \gamma \tag{5}$$

We see now that the model has only $\ell + 1$ genuine degrees of freedom; or, since γ is trivial to read off from $\boldsymbol{\mu}_1$, that the model conditioned on γ has only ℓ degrees of freedom.

If $\boldsymbol{\mu}_1 = 0$ the model is trivial, so from here on out we assume $\gamma > 0$.

To solve for α_{j0} and α_{j1}, up to the symmetry between them, it suffices, in view of (5), to solve for

$$a_j \coloneqq \alpha_{j0}\alpha_{j1}.$$

This transformation results in a family of polynomials (parameterized by γ) q_t in $a_1, \ldots, a_\ell$ such that for any $t \geq 1$, $\boldsymbol{\mu}_t = q_t(a_1, \ldots, a_\ell)$; fixing any m, the mapping $(q_2, \ldots, q_m)$ of $\{a_1, \ldots, a_\ell\}$ to $(\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_m$ carries $\mathbb{R}^\ell$ into a variety of dimension at most ℓ. If the dimension is ℓ, then for all but a set of measure 0, specifically at all regular points of the mapping, the image has finitely many pre-images.

Our main result in this section is:

Theorem 1. *The mapping $(\gamma, a_1, \ldots, a_\ell)$ to $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\ell+1})$ is a.e. locally identifiable.*

In the remainder of the paper, we have relegated the proofs of some technical lemmas to the supplementary material, and replaced lengthy proofs of important results with proof sketches.

2.2 The polynomial sequence

We require a more detailed understanding of the probabilities $\boldsymbol{\mu}_t$. Observe that

$$\alpha_{j0}^2 + \alpha_{j1}^2 = (\alpha_{j0} + \alpha_{j1})^2 - 2\alpha_{j0}\alpha_{j1} = \gamma^2 - 2a_j$$

so that

$$\boldsymbol{\mu}_2 = \prod_{j=1}^{\ell} \frac{\gamma^2 - 2a_j}{2}.$$

Similarly:

$$\begin{aligned}\alpha_{j0}^3 + \alpha_{j1}^3 &= (\alpha_{j0} + \alpha_{j1})^3 - 3(\alpha_{j0}^2\alpha_{j1} + \alpha_{j0}\alpha_{j1}^2) \\ &= \gamma^3 - 3\gamma a_j\end{aligned}$$

so that $\boldsymbol{\mu}_3$ has an analogous expression. This continues. In the remainder of this Section the subscript 'j' is suppressed.

Lemma 2 (Three-term recurrence)**.** *Letting $a = \alpha_0\alpha_1$, $(\alpha_0^m + \alpha_1^m)/2$ can be written as a polynomial $\boldsymbol{p}_m(a)$ satisfying the three-term recurrence*

$$\begin{aligned}\boldsymbol{p}_0(a) &= 1, \\ \boldsymbol{p}_1(a) &= \gamma/2, \\ \boldsymbol{p}_m(a) &= \gamma\boldsymbol{p}_{m-1} - a\boldsymbol{p}_{m-2}, \quad m \geq 2. \end{aligned} \tag{6}$$

Proof. This is the Newton identity relating power and elementary symmetric functions, specialized to the two-variable case; γ is the first, and a is the second, elementary symmetric function of α_0 and α_1. □

The recurrence (6) resembles the familiar recurrence of orthogonal polynomials, but differs significantly in that the variable (here a) multiplies $\boldsymbol{p}_{m-2}$ rather than $\boldsymbol{p}_{m-1}$. In particular the polynomials $\boldsymbol{p}_m$ are not of incrementing degree in a.

(We note however that at the particular value $a = 1/4$, these polynomials as functions of γ are rescalings of the Chebychev polynomials of the first kind.)

Observation 3. *The polynomials $\boldsymbol{p}_m$ defined above satisfy the following:*

1. *The degree of $\boldsymbol{p}_m(a)$ is $\lfloor m/2 \rfloor$.*
2. *$\boldsymbol{p}_m(0) = \gamma^m/2$ for $m \geq 1$.*
3. *The leading term of $\boldsymbol{p}_m(a)$ has a negative sign if $\lfloor m/2 \rfloor$ is odd, and a positive sign if $\lfloor m/2 \rfloor$ is even.*

Proposition 4 (Interlacing)**.** *$\boldsymbol{p}_0$ and $\boldsymbol{p}_1$ are positive constants. For any $m \geq 2$:*

1. *The roots of $\boldsymbol{p}_m$ are simple and contained in the interval $(0, \infty)$; denote them $\beta_{m,1} < \ldots < \beta_{m,\lfloor m/2 \rfloor}$.*
2. *$0 < \beta_{m,1} < \beta_{m-1,1}$ and $\beta_{m-1,i-1} < \beta_{m,i} < \beta_{m-1,i}$ for $i = 2, \ldots, \lfloor m/2 \rfloor$. If $\boldsymbol{p}_m$ has degree greater than $\boldsymbol{p}_{m-1}$ then $\beta_{m-1,\lfloor (m-1)/2 \rfloor} < \beta_{m,\lfloor m/2 \rfloor}$.*
   (For $m = 2$ this requires the convention $\beta_{1,0} = 0, \beta_{1,1} = \infty$.)

Proof sketch. By induction. Let m even for simplicity. By the three term recurrence, $\boldsymbol{p}_m(\beta_{m-1,i}) = -\beta_{m-1,i}\boldsymbol{p}_{m-2}(\beta_{m-1,i})$. By induction, this sign alternates as i ranges from 1 to $m/2 - 1$, giving $m/2 - 2$ roots by the Int. Val. Theorem (IVT). Moreover, $\boldsymbol{p}_m(0) = \gamma^m/2 > 0$, and $\boldsymbol{p}_m(\beta_{m-1,1}) = -\beta_{m-1,1}\boldsymbol{p}_{m-2}(\beta_{m-1,1}) < 0$ because $\beta_{m-1,1} < \beta_{m-2,1}$, accounting for another root. The last root can be found by observing that the sign of the leading term of $\boldsymbol{p}_m$ is opposite of $\boldsymbol{p}_{m-1}, \boldsymbol{p}_{m-2}$ and the IVT. Full proof deferred to Appendix A.1. □

2.3 Identifiability of latent symmetric models: the method

We now describe the algebraic tool which enables the proof of Theorem 1. Consider a sequence of ℓ univariate polynomials $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_\ell$. (We will eventually substitute the $\boldsymbol{P}$'s of the previous Section.) Let $y_1, \ldots, y_\ell$ be indeterminates, and $y = (y_1, \ldots, y_\ell)$. Construct symmetric polynomials in the y_j by taking products as follows:

$$q_m(y) = \prod_{j=1}^{\ell} \boldsymbol{P}_m(y_j).$$

Proposition 5. *Suppose that for every $m \in [\ell]$, $\boldsymbol{P}_m$ has a root η_m that is simple and is not a root of $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_{m-1}$. Then the mapping M_ℓ where*

$$(y_1, \ldots, y_\ell) \mapsto (q_1(y), \ldots, q_\ell(y))$$

is locally identifiable.

Proof sketch. We aim to construct a full-rank diagonal Jacobian. If a is the indeterminate for $\boldsymbol{P}_i$, then

$$\frac{\partial q_i}{\partial y_j} = \frac{\partial \boldsymbol{P}_i}{\partial a}(\eta_j) \prod_{k \neq j} \boldsymbol{P}_i(\eta_k).$$

Note that $\partial \boldsymbol{P}_i / \partial a(\eta_i) \neq 0$ by simplicity of roots and $\boldsymbol{P}_i(\eta_i) = 0$, giving us a mechanism to force off-diagonal entries to be zero and hope that on-diagonal entries are nonzero. By composing M_ℓ with another function, we are able to achieve this exactly. Full proof deferred to Appendix A.2. A proof of a more general version of this proposition can be found in Appendix A.3. □

Proof of Theorem 1. Since γ is identified from $\boldsymbol{\mu}_1$, we have only to show that for any fixed $\gamma > 0$, the mapping $\{a_1, \ldots, a_\ell\} \mapsto (\boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{\ell+1})$ is locally identifiable.

To show this we apply Prop. 5 with the polynomials $\boldsymbol{P}_m = \boldsymbol{p}_{m+1}$ for $m = 1, \ldots, \ell$ (for $\boldsymbol{p}_m$ as defined in Lemma 2), and the root η_m of $\boldsymbol{P}_m$ being the point $\beta_{m+1,1}$ provided by Prop. 4. □

3 IDENTIFYING MODELS WITH GENERAL PRIORS

3.1 Preliminaries

We now treat the more general setting where U_j have arbitrary priors, specified by the parameters

$$\pi_j = \Pr(U_j = 1).$$

Because the U_j's are not uniformly sampled, the moments will take on a different form than before (we call the nth moment q_n):

$$\begin{aligned} q_n &\coloneqq \Pr(X_1 = \cdots = X_n = 1) \\ &= \sum_{u \in \{0,1\}^\ell} \prod_{j=1}^{\ell} \pi_j(u_j) \alpha_{ju_j}^n \\ &= \prod_{j=1}^{\ell} \left((1-\pi_j)\alpha_{j0}^n + \pi_j \alpha_{j1}^n \right) \\ &= \prod_{j=1}^{\ell} r_n(\alpha_{j0}, \alpha_{j1}, \pi_j) \end{aligned} \tag{7}$$

where we define $r_n(\alpha_0, \alpha_1, \pi) = \pi\alpha_1^n + (1-\pi)\alpha_0^n$. We may view $r_n(\alpha_{j0}, \alpha_{j1}, \pi_j)$ as the contribution of U_j to the moment q_n.

Symmetries of the Model The model has discrete and continuous symmetries as before. The moments q_t are invariant to:

1. *Discrete symmetries (hyperoctahedral).*
   (a) For any j, exchange α_{j0} and α_{j1}, and replace π_j with $1 - \pi_j$.
   (b) Exchange any j and j'. That is, for $j \neq j'$, exchange α_{j0} with $\alpha_{j'0}$, α_{j1} with $\alpha_{j'1}$, and π_j with $\pi_{j'}$.
2. *Continuous symmetries.* For any $j \neq j'$ and $\lambda > 0$, the gauge transformation

$$\begin{cases} (\alpha_{j0}, \alpha_{j1}) & \mapsto (\lambda\alpha_{j0}, \lambda\alpha_{j1}), \\ (\alpha_{j'0}, \alpha_{j'1}) & \mapsto (\lambda^{-1}\alpha_{j'0}, \lambda^{-1}\alpha_{j'1}) \end{cases}.$$

Of course, the model may only be identified up to these symmetries. Therefore, as before, we use the gauge symmetries to scale our parameters such that for all $j \in [\ell]$, letting $\gamma \coloneqq \Pr(X_1 = 1)^{1/\ell}$, we have that $r_1(\alpha_{j0}, \alpha_{j1}, \pi_j) = \gamma$. As in Sec. 2, the model is trivial if $\gamma = 0$, so we assume throughout that $\gamma \neq 0$.

3.2 Polynomial sequences

It will be convenient to make the change of variables $\sigma_j = 2\pi_j - 1$. Due to the factorization (7), we can while studying the polynomials r, focus on an arbitrary j, and drop the indices j until we return to treating the polynomials q. We can now write

$$r_n = \pi\alpha_1^n + (1-\pi)\alpha_0^n = \frac{\alpha_1^n + \alpha_0^n}{2} + \sigma\frac{\alpha_1^n - \alpha_0^n}{2}.$$

Let $d = (\alpha_1 - \alpha_0)/2$, expanding r_n for $n > 1$ we get:

$$r_n = 2(\gamma - \sigma d) r_{n-1} - [(\gamma - \sigma d)^2 - d^2] r_{n-2} \tag{8}$$

The derivation for this recurrence can be found in Appendix A.4. We make a final change of variables to replace (σ, d) by (x, y): $x = \gamma - \sigma d$ and $y = d^2$. (This is invertible after quotienting by the hyperoctahedral symmetry of the model.) This yields the following three-term recurrence:

$$\begin{aligned} r_0(x, y) &= 1 \\ r_1(x, y) &= \gamma \\ r_n(x, y) &= 2x r_{n-1} - (x^2 - y) r_{n-2}. \end{aligned}$$

It is helpful to define the following family of polynomials (p_n), which are very closely related to the polynomials (r_n). In particular, we will see in the following proposition that they are the "coefficient polynomials" of the r_n.

$$\begin{aligned} p_{-1}(x, y) &= 1 \\ p_0(x, y) &= 2x \\ p_n(x, y) &= 2x p_{n-1} - (x^2 - y) p_{n-2} \end{aligned}$$

Proposition 6. *For any* $0 \le k \le n-1$,

$$r_n = p_k r_{n-k-1} - (x^2 - y) p_{k-1} r_{n-k-2}$$
$$p_n = p_k p_{n-k-1} - (x^2 - y) p_{k-1} p_{n-k-2}$$

Proof. By induction over k. Full proof deferred to Appendix A.5. □

In analogy to Section 2.2, where we studied the roots of the univariate polynomials $\boldsymbol{p}_n$, we now need some understanding of where each r_n (which is bivariate, in variables x, y) is zero.

Notice that $r_i(0,0) = 0$ for all $i \ge 2$. Since this is a common zero for all r_i, $i \ge 2$, we call this the trivial zero. The proof of the following statement is in the supplementary material.

Lemma 7. *For $i \ge 2$ the only zero of r_i on the curves $x = 0$ and $x^2 = y$ is the trivial zero.*

Proof deferred to Appendix A.6.

3.3 Common Zeros

A new phenomenon that we encounter, unlike in Section 2, is that we need to identify *common zeros* of r_i, r_j for $i \ne j$. First we make the following observation.

Lemma 8. *For no i is there a nontrivial zero shared by r_i and r_{i+1}, or by r_i and r_{i+2}.*

Proof. Pick the smallest such i for which either claim fails, and let (x_0, y_0) be a nontrivial zero. We know from Lemma 7 that $x_0 \ne 0$ and $x_0^2 \ne y_0$. If the claim fails because $r_i = r_{i+2} = 0$, then writing $r_{i+2} = 2x_0 r_{i+1} - (x_0^2 - y_0) r_i$, we see that necessarily also $r_{i+1} = 0$. Then

$$r_{i+1} = 2x_0 r_i - (x_0^2 - y_0) r_{i-1}$$
$$0 = (x_0^2 - y_0) r_{i-1}$$

Since $x_0^2 \ne y_0$, it follows that $r_{i-1} = 0$, which contradicts the minimality of i. □

The structure of pairwise-common roots in the (x, y) plane is complex and has two especially interesting regions: the line $x = \gamma/2$ and the parabola $y = x^2 - \gamma^2$. The remainder of our analysis relies on common roots within the first of these regions.

3.4 Restricting to the line $x = \gamma/2$

On this line we have the recurrence:

$$r_n(\gamma/2, y) = \gamma r_{n-1} - (\gamma^2/4 - y) r_{n-2} \qquad (9)$$

Since $p_0(\gamma/2, y) = \gamma$, the initial conditions for the p polynomials and r polynomials are identical evaluated on the line:

$$p_{-1}(\gamma/2, y) = r_0(\gamma/2, y)$$
$$p_0(\gamma/2, y) = r_1(\gamma/2, y)$$

Observe that both p and r polynomials have the same recurrence on the line, so it must be true that $r_n(\gamma/2, y) = p_{n-1}(\gamma/2, y)$. Now consider the univariate polynomials defined by the following recursion:

$$s_0(y) = 0$$
$$s_1(y) = 1$$
$$s_n(y) = \gamma s_{n-1} - (\gamma^2/4 - y) s_{n-2}$$

Notice that $r_{n-1}(\gamma/2, y) = p_{n-2}(\gamma/2, y) = s_n(y)$, thus we will turn our attention to the zeros of the s polynomials. From Proposition 6 we have the following corollary:

Corollary 9. *For all $2 \le k \le n-1$, $s_n = s_k s_{n-k+1} - (\gamma^2/4 - y) s_{k-1} s_{n-k}$.*

We can now prove the following, which has no analogue in the uniform-priors case but is key to the general-priors case.

Theorem 10. $\gcd(s_i, s_j) = s_{\gcd(i,j)}$.

(Note, this is gcd in the ring $\mathbb{R}[y]$.)

Proof. The theorem will follow from showing that:

$$\text{If } j > i \text{ then } \gcd(s_i, s_j) = \gcd(s_i, s_{j-i}). \qquad (10)$$

To show (10): In Corollary 9, since $j > i$, we can substitute $n = j$ and $k = i + 1$ to obtain:

$$s_j = s_{i+1} s_{j-i} - (\gamma^2/4 - y) s_i s_{j-i-1}$$

This shows $\gcd(s_i, s_{j-i}) \mid s_j$ therefore $\gcd(s_i, s_{j-i}) \mid \gcd(s_i, s_j)$. This also shows $\gcd(s_i, s_j) \mid s_{i+1} s_{j-i}$. From Lemma 8, s_i and s_{i+1} are relatively prime, it follows that $\gcd(s_i, s_j) \mid s_{j-i}$. Consequently $\gcd(s_i, s_j) \mid \gcd(s_i, s_{j-i})$. □

3.5 Simple roots of s_n polynomials

We aim to show here that the roots of s_n are simple and real. We start with some useful properties. Recall that $\gamma \in (0, 1]$.

Lemma 11. *For every $n \ge 1$,*

1. *The leading coefficient of s_n is positive.*
2. $s_n(0) = n(\frac{\gamma}{2})^{n-1} > 0$.

3. *The degree of* s_n *is* $\lfloor \frac{n-1}{2} \rfloor$.

Proof deferred to Appendix A.7. Using Lemma 11 we exhibit an interlacing property of the polynomials s_n, with similar proof to Prop. 4.

Lemma 12. *For* $k \geq 3$, s_k *has real roots* $\beta_{k,1}, \ldots, \beta_{k,\lfloor (k-1)/2 \rfloor}$, *satisfying the following: (a) For* $n > 1$, $\beta_{2n+1,1} < \beta_{2n,1} < \cdots < \beta_{2n,n-1} < \beta_{2n+1,n} < 0$. *(b) For* $n \geq 1$, $-\infty < \beta_{2n+1,1} < \beta_{2n+2,1} < \cdots < \beta_{2n+1,n} < \beta_{2n+2,n} < 0$.

In particular, each s_k *has only simple roots.*

Proof deferred to Appendix A.8.

3.6 Decomposition of the polynomials s_n

Theorem 10 and Lemma 12 actually imply that the polynomials s_n have considerably more structure than already revealed. This will be essential to our results.

Lemma 13. *There are real polynomials* h_n *such that* h_n *has only simple real roots,* $\gcd(h_n, h_m) = 1$ *for* $n \neq m$, *and for every* n,

$$s_n(y) = \prod_{d|n} h_d(y). \tag{11}$$

Proof. We prove this by induction on n. For the gcd claim, while treating n we address only m s.t. $n \nmid m$.

Since $s_1 = 1$, the factorization and real-roots claims hold for n prime, with $h_n = s_n$; the gcd claim follows from Theorem 10.

For n composite, we know from Theorem 10 that s_n shares the roots of every s_d, $d \mid n$; by the inductive hypothesis this is equivalent to saying that s_n is divisible by $\prod_{d|n, d<n} h_d(y)$. Since s_n has only simple roots, any remaining factors of s_n cannot be shared with any h_d for $d \mid n, d < n$. Set $h_n = s_n / \left(\prod_{d|n,d<n} h_d(y)\right)$. Again, since s_n has only simple roots, h_n is relatively prime to every $h_d, d \mid n$. It remains to show that h_n is relatively prime to h_m for $n \nmid m$. Since $\gcd(s_n, s_m) = s_{\gcd(n,m)}$, which does not include any factors of h_m, we know that h_m is relatively prime to s_n, and therefore to h_n. □

We will refer to the h_n as "atomic polynomials" to recognize that they are what compose the s_n polynomials. Next we wish to work out their degrees, which we denote $f(n) = \deg h_n$.

Lemma 14. $f(1) = 0$, $f(2) = 0$, *and for* $n > 2$, $f(n) = \frac{n}{2} \prod_{q \text{ prime}, q|n} (1 - 1/q) > 0$.

Proof deferred to Appendix A.9.

For every $n > 2$, therefore, h_n possesses a nonempty set of simple roots, called the atomic roots of h_n or s_n; these are roots of s_m if and only if $n \mid m$. See Fig. 2.

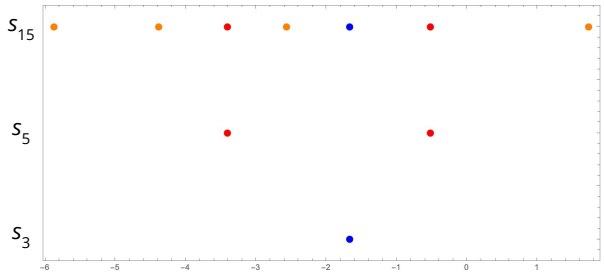

Figure 2: The roots of s_3, s_5, and s_{15}. The blue, red and orange dots represent the roots of the h_3, h_5 and h_{15} atomic polynomials respectively. For better spacing, the transformation $x \mapsto \log(-x)$ was applied to the horizontal axis.

3.7 Identifiability: the Jacobian at perturbed common roots

In this section we finally show how to generalize the method of Sec. 2.3, by leveraging roots shared between pairs of s_n's. Unlike in the basic method introduced in Sec. 2.3, it will not work to evaluate the Jacobian of the mapping exactly at a point specified by these common roots; instead, it will be necessary to slightly perturb the evaluation point.

For any pair r_i and r_{2i+1}, let J_i be the Jacobian of the map $(x, y) \mapsto (r_i(x, y), r_{2i+1}(x, y))$. Proposition 6 tells us $r_{2i+1} = p_{i-1} r_{i+1} - (x^2 - y) p_{i-2} r_i$, we compute the partial derivatives of r_{2i+1} using this expression and the determinant of J_i simplifies to:

$$r_{i+1} \left(\frac{\partial r_i}{\partial x} \frac{\partial p_{i-1}}{\partial y} - \frac{\partial r_i}{\partial y} \frac{\partial p_{i-1}}{\partial x} \right) + p_{i-1} F_i + r_i G_i \tag{12}$$

$$F_i = \frac{\partial r_i}{\partial x} \frac{\partial r_{i+1}}{\partial y} - \frac{\partial r_i}{\partial y} \frac{\partial r_{i+1}}{\partial x} \tag{13}$$

$$G_i = p_{i-2} \left(\frac{\partial r_i}{\partial x} + 2x \frac{\partial r_i}{\partial y} \right) - (x^2 - y) \left(\frac{\partial r_i}{\partial x} \frac{\partial p_{i-2}}{\partial y} - \frac{\partial r_i}{\partial y} \frac{\partial p_{i-2}}{\partial x} \right) \tag{14, 15}$$

The proof of the following Lemma follows from using Proposition 6 to write both r_i and p_{i-1} in terms of p_{i-2} and p_{i-3} then simply plugging in $x = \gamma/2$ into Equation 12. The proof in its entirety is in the supplementary material.

Lemma 15. *On the line* $x = \gamma/2$ *the determinant of* J_i *is* $-2 s_{i+2} s_i \frac{\partial s_{i+1}}{\partial y} + s_{i+1}(F_i + G_i)$

Proof deferred to Appendix A.10.

Now we state the main theorem of this section.

Theorem 16. *The map* $(x_1, y_1, \ldots, x_\ell, y_\ell) \mapsto (q_2, q_5, \ldots, q_{2n}, q_{4n+1}, \ldots, q_{2\ell}, q_{4\ell+1})$ *is locally identifiable.*

The proof of this statement is lengthy and we have deferred it to Appendix A.11, but we will show the proof of a simpler case. Let p_i be the ith prime. Consider the following map:

$$(x_1, \ldots, y_\ell) \mapsto (q_{p_2-1}, q_{2p_2-1}, \ldots, q_{p_{\ell+1}-1}, q_{2p_{\ell+1}-1}) \tag{16}$$

We can explicitly write down the entry in the ith row and the jth column of the Jacobian evaluated at $(x_1, y_1, \ldots, x_\ell, y_\ell)$. Suppose $i = 2n-1$ and $j = 2m-1$:

$$\frac{\partial r_{p_{n+1}-1}}{\partial x}(x_n, y_n) \cdot \prod_{k \neq n} r_{p_{n+1}}(x_k, y_k) \tag{17}$$

If $i = 2n$, the term $r_{p_{n+1}-1}$ is replaced with $r_{2p_{n+1}-1}$ and when $j = 2m$ the partial is taken with respect to y instead. This allows us to look at the determinant in 2×2 blocks.

Now pick the point to evaluate the Jacobian as follows. Let c_i be a root of $s_{p_{i+1}}$ which exists since $p_{i+1} > 2$ so $s_{p_{i+1}}$ is non-constant. From Theorem 10 we know that this is also a root of $s_{2p_{i+1}}$, and so $(\gamma/2, c_i)$ is a root of both $r_{p_{i+1}-1}$ and $r_{2p_{i+1}-1}$. Recall all roots of s polynomials are simple so for all i:

$$\frac{\partial s_{p_{i+1}}}{\partial y}(\gamma/2, c_i) \neq 0 \qquad \frac{\partial s_{2p_{i+1}}}{\partial y}(\gamma/2, c_i) \neq 0 \tag{18}$$

It follows from Equation 17 that all entries not on a diagonal block of the Jacobian evaluated at $(\gamma/2, c_1, \ldots, \gamma/2, c_\ell)$ are zero. This makes the Jacobian block diagonal; the determinant of each diagonal block is:

$$\left(\prod_{k \neq i} s_{2p_{i+1}}(c_k) \cdot s_{p_{i+1}}(c_k) \right) \cdot \det(J_{p_{i+1}-1}(\gamma/2, c_i))$$

From Theorem 10, for $i \neq j \in [\ell]$ in the ring $\mathbb{R}[y]$ we see $\gcd(s_{2p_i}, s_{p_j}) = \gcd(s_{p_i}, s_{p_j}) = s_1 = 1$ and $\gcd(s_{2p_i}, s_{2p_j}) = s_2 = \gamma$ so c_k is only a root of $s_{p_{k+1}}$ and $s_{2p_{k+1}}$. Applying Lemma 15 and using Equation 18 we see that the determinant of each diagonal block is non-zero. Since the determinant of the Jacobian is the product over the determinant of each diagnoal block, it is non-zero and the map defined by Equation 16 is locally identifiable.

Unfortunately, this method cannot be generalized to prove Theorem 16 because the polynomials s_{2i+1} and s_{2j+1} are not relatively prime in $\mathbb{R}[y]$. However, by ϵ-perturbing atomic roots, we can bound the order of each entry of the Jacobian. By scaling the Jacobian evaluated at these perturbed roots, all of the entries in 2×2 blocks above the diagonal scale with some small ϵ. It is well known that for a block lower-triangular matrix, the determinant is the product of its diagonal blocks because any permutation that selects an entry not in the diagonal block is either zero or will force the permutation to select a zero in the other half. Analogously, we show that the determinant of the product of the diagonal blocks contains a term independent of ϵ and all other terms contributed to the determinant scale with ϵ. We then pick ϵ sufficiently small so that the matrix is non-singular, showing local identifiability.

4 CORRESPONDENCE BETWEEN THE LATENT VARIABLE FRAMEWORK AND THE STANDARD PoE FRAMEWORK

Here, we justify the correspondence between the model discussed in this paper and the traditional formulation of the PoE. The PoE is often discussed in the literature; we will use [Oneto and Vannieuwenhoven, 2023] as a recent reference point. They define the PoE's distribution by a normalized Hadamard product of weighted rank 1 tensors. In comparison, in our model, the observable statistics are precisely the set of probabilities that a given subset of the m observable bits equal 1, and the rest equal 0. In other words, if we let $X = (X_1, X_2, \ldots, X_m)$, then the observable statistics are $\Pr(X = s)$ for $s \in \{0, 1\}^m$. Since in our model the X_i are i.i.d. conditional on the latents, and hence exchangeable, $\Pr(X = s) = \Pr(X = s')$ if $w(s) = w(s')$ where w is the Hamming weight, so the statistics $\Pr(X_1 = \cdots = X_t = 1)$, for $t = 1, \ldots, m$ are sufficient to describe the full set of statistics. To see the correspondence with the standard PoE formulation, note that we may write our statistics in the form of the $2 \times \cdots \times 2$ (m times) tensor ($\odot$ is Hadamard product):

$$M = \bigodot_{j=1}^{\ell} \left(\pi_j \begin{bmatrix} 1 \\ \alpha_{j1} \end{bmatrix}^{\otimes m} + (1 - \pi_j) \begin{bmatrix} 1 \\ \alpha_{j0} \end{bmatrix}^{\otimes m} \right) \tag{19}$$

Compare this with Eqn. (1.3) in [Oneto and Vannieuwenhoven, 2023] (using our index ℓ):

$$\mathscr{P} = \lambda \cdot \mathscr{P}^{(1)} \odot \cdots \odot \mathscr{P}^{(\ell)} \tag{20}$$

where λ is a normalizing (partition function) factor.

Thus M and $\mathscr{P}$ have precisely the same form, except that in our case, we are missing the normalization factor λ, because the entries of M do not themselves add up to a probability distribution; rather they are the probabilities of certain events. Specifically, if $w(s) = t$ for $s \in \{0,1\}^m$, then from Eqn. (19):

$$M[s] = \prod_{j=1}^{\ell} \left((1-\pi_j)\alpha_{j0}^t + \pi_j \alpha_{j1}^t\right) = \Pr(X_1 = \cdots X_t = 1).$$

However, in both our formulation and the standard PoE formulation Eqn. (20), the identification problem is that of producing the decomposition of a tensor of precisely the same type.

5 DISCUSSION

Related work: undirected graphical models. The *conditional distributions* on X that occur in our model (1) agree with those of Restricted Boltzmann Machines (RBM) [Ackley et al., 1985] (a kind of Markov random field model), and particularly the "harmonium" [Smolensky, 1986, Freund and Haussler, 1991] special case. An RBM is an undirected graphical model comprising one layer of latent random variables $\mathcal{U}$, one layer of observable random variables $\mathcal{X}$, and satisfying that the X_i are independent conditional on $\mathcal{U}$. An RBM is often written in the following form:

$$\Pr((\mathcal{X},\mathcal{U}) = (x,u)) = \frac{1}{Z}\exp(-xWu^\dagger - xb^\dagger - cu^\dagger) \tag{21}$$

where $Z = \sum_{x,u} \exp(-xWu^\dagger - xb^\dagger - cu^\dagger)$. The dependence of X on U can be expressed:

$$\Pr(x|u) = \exp(-xWu^\dagger - xb^\dagger - cu^\dagger - d) \tag{22}$$

where $d = -\sum_{i,j}\log(1-\alpha_{i,j,0})$; $b_i = -\sum_j \log\frac{\alpha_{i,j,0}}{1-\alpha_{i,j,0}}$; $c_j = -\sum_i \log\frac{1-\alpha_{i,j,1}}{1-\alpha_{i,j,0}}$; $W_{ij} = -\log\frac{\alpha_{i,j,1}(1-\alpha_{i,j,0})}{\alpha_{i,j,0}(1-\alpha_{i,j,1})}$. This is referred to as an undirected graphical model because each coefficient W_{ij} is regarded as an energy associated with the unordered pair of sites $\{i,j\}$ (i observable, j latent). The conditional distributions $\Pr(\mathcal{X}|\mathcal{U})$ in (22) have the same form as in our model (1) (or its more general version in which the X_i are only conditionally independent)—but (22) does not allow imposition of a chosen product distribution on $\mathcal{U}$ as the prior, and in fact, generally the U_j will not be independent in the distribution (21).

The conditional distributions of RBMs enable expressive statistical models with relatively few parameters. For this reason and because of the connection to layered networks, RBMs have been extensively studied in the neural networks and algebraic statistics literatures (citations above). An interesting recent (and independent) contribution in this literature concerns identifiability [Oneto and Vannieuwenhoven, 2023], but there are no obvious implications in either direction between the works. (The main thing to note is that it relies for identifiability on the Kruskal condition. That condition does gives an upper bound on the number of observables required but if one works out the bound, one sees that it cannot be less than $2^{\ell+1}-1$. But the slightly better bound of 2^ℓ was the exponentially-large bound that we set out to rectify. Second, one should note that that work addresses a somewhat more general class of models, but then relies upon a general-position assumption about the input, an assumption we do not make, and that is also not valid in our setting of conditionally-iid X_i.)

Open questions A fundamental issue is whether there is an algorithm for *efficiently* identifying the model from its statistics. Settling this in the positive would be the ideal way also of proving full identifiability.

A natural question is whether the product in Eqn. (3) can be replaced by other symmetric functions; even more generally, one may consider the situation in which the effect of the latent variables $U_1,\ldots,U_\ell$ on the observable variables is invariant not under the permutation group S_ℓ but under, say, a transitive subgroup of S_ℓ.

A full understanding of this family of problems requires also extending to non-binary X_i and U_j. The lead of [Fan and Li, 2022] may be useful toward the case of non-binary X_i. It appears more challenging to address non-binary U_j, as this demands replacing our two-dimensional space (x,y) by a higher-dimensional space and, perhaps, generalizing our approach through pairwise-common zeros, to zeros shared by larger assemblies of polynomials.

Acknowledgements

The authors are with the Division of Engineering and Applied Science, California Institute of Technology, Pasadena CA 91125 USA (emails: slgordon, mkant, ema, astaicu, schulman@caltech.edu). Research supported in part by NSF grants CCF-1909972 and CCF-2321079; and by the Caltech SURF program and the Larson, Mike Stefanko, and SURF Board endowments. We thank Caroline Uhler, Bernd Sturmfels, Yulia Alexandr and Guido Montúfar for helpful discussions.

References

[Ackley et al., 1985] Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169.

[Blischke, 1964] Blischke, W. R. (1964). Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, 59(306):510–528.

[Cueto et al., 2010] Cueto, M. A., Morton, J., and Sturmfels, B. (2010). Geometry of the restricted Boltzmann machine. In Viana, M. A. G. and Wynn, H. P., editors, *Algebraic Methods in Statistics and Probability II*, volume 516 of *Contemporary Mathematics*, pages 135–153.

[de Prony, 1795] de Prony, R. (1795). Essai expérimentale et analytique. *J. Écol. Polytech.*, 1(2):24–76.

[Drton et al., 2007] Drton, M., Sturmfels, B., and Sullivant, S. (2007). Algebraic factor analysis: tetrads, pentads and beyond. *Probab. Theory Relat. Fields*, 138:463–493.

[Fan and Li, 2022] Fan, Z. and Li, J. (2022). Efficient algorithms for sparse moment problems without separation.

[Freund and Haussler, 1991] Freund, Y. and Haussler, D. (1991). Unsupervised learning of distributions on binary vectors using two layer networks. In Moody, J., Hanson, S., and Lippmann, R. P., editors, *Proc. 4th Int'l Conf. on Neural Information Processing Systems*, pages 912–919. Morgan-Kaufmann.

[Gordon et al., 2020] Gordon, S., Mazaheri, B., Schulman, L. J., and Rabani, Y. (2020). The sparse Hausdorff moment problem, with application to topic models.

[Gordon et al., 2023] Gordon, S. L., Mazaheri, B., Rabani, Y., and Schulman, L. (2023). Causal inference despite limited global confounding via mixture models. In van der Schaar, M., Zhang, C., and Janzing, D., editors, *Proc. Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 574–601. PMLR.

[Gordon et al., 2021] Gordon, S. L., Mazaheri, B., Rabani, Y., and Schulman, L. J. (2021). Source identification for mixtures of product distributions. In *Proc. 34th Ann. Conf. on Learning Theory - COLT*, volume 134 of *Proc. Machine Learning Research*, pages 2193–2216. PMLR.

[Hinton, 1999] Hinton, G. E. (1999). Products of experts. In *Proc. 9th Int'l Conf. on Artificial Neural Networks*, volume 1, pages 1–6.

[Hinton, 2002] Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.

[Hinton, 2010] Hinton, G. E. (2010). A practical guide to training restricted Boltzmann machines, version 1 (UTML2010-003). Technical report, U Toronto.

[Hinton et al., 2006] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554.

[Huang et al., 2022] Huang, X., Mallya, A., Wang, T., and Liu, M. (2022). Multimodal conditional image synthesis with product-of-experts GANs. *European Conference on Computer Vision*, 13676:91–109.

[Kim et al., 2019] Kim, Y., Koehler, F., Moitra, A., Mossel, E., and Ramnarayan, G. (2019). How many subpopulations is too many? Exponential lower bounds for inferring population histories. In Cowen, L., editor, *Int'l Conf. on Research in Computational Molecular Biology*, volume 11457 of *Lecture Notes in Computer Science*, pages 136–157. Springer.

[Liu et al., 2021] Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., and Choi, Y. (2021). DExperts: Decoding-time controlled text generation with experts and anti-experts. *Association for Computational Linguistics*.

[Long and Servedio, 2010] Long, P. M. and Servedio, R. A. (2010). Restricted Boltzmann machines are hard to approximately evaluate or simulate. In *Proc. 27th Int'l Conf. on Machine Learning*, ICML, page 703–710. Omnipress.

[Martens et al., 2013] Martens, J., Chattopadhyay, A., Pitassi, T., and Zemel, R. (2013). On the representational efficiency of restricted Boltzmann machines. In *Proc. Neurips*, volume 26, pages 2877–2885. Curran Associates.

[Montúfar, 2018] Montúfar, G. (2018). Restricted Boltzmann machines: Introduction and review.

[Montúfar and Morton, 2015] Montúfar, G. and Morton, J. (2015). Discrete restricted Boltzmann machines. *J. Machine Learning Research*, 16(21):653–672.

[Montúfar and Morton, 2017] Montúfar, G. and Morton, J. (2017). Dimension of marginals of Kronecker product models. *SIAM J. Appl. Algebra Geometry*, 1(1):126–151.

[Oneto and Vannieuwenhoven, 2023] Oneto, A. and Vannieuwenhoven, N. (2023). Hadamard-Hitchcock decompositions: identifiability and computation.

[Roux and Bengio, 2008] Roux, N. L. and Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649.

[Seigal and Montúfar, 2018] Seigal, A. and Montúfar, G. (2018). Mixtures and products in two graphical models. *J. Algebr. Stat.*, 9(1):1–20.

[Smolensky, 1986] Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, chapter 6, pages 194–281. MIT Press, Cambridge, MA, USA.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Section 1, Products of Experts Models, Section 2.1, and Section 3.1
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] Some proofs are in the supplemental material because they do not fit in the main paper
   (c) Clear explanations of any assumptions. [Yes] See Section 1 Products of Experts Models, Section 2 Symmetries of the Model, and Section 3 Symmetries of the Model

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A DEFERRED PROOFS

A.1 Proof of Proposition 4

Proof. We induct on m. The Proposition holds for $m = 0, 1$. Now $\boldsymbol{p}_2 = -a + \gamma^2/2$ which has a root at $\gamma^2/2$ and $\boldsymbol{p}_3 = \gamma(-a + \gamma^2/2) - a\gamma/2 = -3\gamma a/2 + \gamma^3/2$ which has a root at $\gamma^2/3$.

Now fix any $m \geq 2$. Let $d = \deg(\boldsymbol{p}_{m-2})$, so $d + 1 = \deg(\boldsymbol{p}_m)$. By the inductive hypothesis, $0 < \beta_{m-1,1} < \beta_{m-2,1} < \beta_{m-1,2} < \cdots < \beta_{m-1,d} < \beta_{m-2,d}$. If $\deg(\boldsymbol{p}_{m-1}) = d + 1 > \deg(\boldsymbol{p}_m)$, then there is an additional root $\beta_{m-1,d+1}$ of $\boldsymbol{p}_{m-1}$ with $\beta_{m-2,d} < \beta_{m-1,d+1}$. Since we've accounted for every root of $\boldsymbol{p}_{m-1}$ and $\boldsymbol{p}_{m-2}$, the value of $\boldsymbol{p}_{m-1}$ must alternate between strictly positive and strictly negative on the sequence of open intervals $(-\infty, \beta_{m-1,1})$, $(\beta_{m-1,1}, \beta_{m-1,2})$, $(\beta_{m-1,2}, \beta_{m-1,3}), \ldots,$ $(\beta_{m-1,\lfloor (m-1)/2 \rfloor}, \infty)$, and $\boldsymbol{p}_{m-2}$ alternates in sign on the intervals $(-\infty, \beta_{m-2,1})$, $(\beta_{m-2,1}, \beta_{m-2,2}), \ldots,$ $(\beta_{m-2,d}, \infty)$. Now we compute

$$\begin{aligned}
\boldsymbol{p}_m(\beta_{m-1,1}) &= \gamma \boldsymbol{p}_{m-1}(\beta_{m-1,1}) - \beta_{m-1,1}\boldsymbol{p}_{m-2}(\beta_{m-1,1}) &&= -\beta_{m-1,1}\boldsymbol{p}_{m-2}(\beta_{m-1,1}) &&< 0; \\
\boldsymbol{p}_m(\beta_{m-2,1}) &= \gamma \boldsymbol{p}_{m-1}(\beta_{m-2,1}) - \beta_{m-2,1}\boldsymbol{p}_{m-2}(\beta_{m-2,1}) &&= \gamma \boldsymbol{p}_{m-1}(\beta_{m-2,1}) &&< 0.
\end{aligned}$$

By Observation 3, $\boldsymbol{p}_m(0) = \gamma^m/2 > 0$ so there must be a root of $\boldsymbol{p}_m$ in $(0, \beta_{m-1,1})$.

For $1 < i < d$, $\beta_{m-1,i} \in (\beta_{m-2,i-1}, \beta_{m-2,i})$ and $\beta_{m-2,i} \in (\beta_{m-1,i}, \beta_{m-1,i+1})$. Moreover, $\boldsymbol{p}_m(\beta_{m-1,i}) = -\beta_{m-1,i}\boldsymbol{p}_{m-2}(\beta_{m-1,i})$ and $\boldsymbol{p}_m(\beta_{m-2,i}) = \gamma \boldsymbol{p}_{m-1}(\beta_{m-2,i})$, so $\operatorname{sign}(\boldsymbol{p}_m(\beta_{m-1,i})) = \operatorname{sign}(\boldsymbol{p}_m(\beta_{m-2,i})) = -\operatorname{sign}(\boldsymbol{p}_m(\beta_{m-1,i+1}))$. We conclude that there is a root of $\boldsymbol{p}_m$ in the interval $(\beta_{m-2,i}, \beta_{m-1,i+1})$ for $1 < i < d$.

We've shown that there are roots of $\boldsymbol{p}_m$ in each of the intervals $(0, \beta_{m-1,1})$, $(\beta_{m-2,1}, \beta_{m-1,2})$, $(\beta_{m-2,2}, \beta_{m-1,3}), \ldots,$ $(\beta_{m-2,d-1}, \beta_{m-1,d})$. If $\deg(\boldsymbol{p}_{m-1}) = d + 1$, then by the same logic there is also a root in $(\beta_{m-2,d}, \beta_{m-1,d+1})$ and the proof is complete. If $\deg(\boldsymbol{p}_{m-1}) = d$, then the leading term of $\boldsymbol{p}_m$ has a different sign than the leading terms of $\boldsymbol{p}_{m-1}$ and $\boldsymbol{p}_{m-2}$. Since $\operatorname{sign}(\boldsymbol{p}_m(\beta_{m-2,d})) = \operatorname{sign}(\boldsymbol{p}_{m-1}(\beta_{m-2,d}))$ and $\beta_{m-2,d}$ is greater than all the roots of $\boldsymbol{p}_{m-1}$ it must be the case that $\operatorname{sign}(\boldsymbol{p}_{m-1}(y)) = \operatorname{sign}(\boldsymbol{p}_{m-1}(\beta_{m-2,d}))$ for all $y \in [\beta_{m-2,d}, \infty)$. But $\lim_{y\to\infty} \boldsymbol{p}_m(y) = -\lim_{y\to\infty} \boldsymbol{p}_{m-1}(y)$, so there must be a root of $\boldsymbol{p}_m$ in $(\beta_{m-2,d}, \infty)$. We've thus accounted for all $d + 1$ roots of $\boldsymbol{p}_m$. □

A.2 Proof of Proposition 5

Proof. It suffices to show that there is a point at which the Jacobian of the mapping is nonsingular. In what follows for a rational function g let $M(g, \eta_j)$ denote the multiplicity of η_j as a root of g; if η_j is a pole of g then $-M(g, \eta_j)$ is the order of the pole.

By assumption, $M(\boldsymbol{p}_i, \eta_j) = 0$ for $j > i$ and $M(\boldsymbol{p}_j, \eta_j) = 1$ for all j.

We now construct a sequence of rational functions $r_1, \ldots, r_\ell$ satisfying (with δ_{ij} = Kronecker delta):

$$M(r_i, \eta_j) = \delta_{ij}.$$

First, we set $r_1 = \boldsymbol{p}_1$, since $M(\boldsymbol{p}_1, \eta_j) = 0$ for all $j > 1$.

Inductively we construct r_i for $i \geq 2$ as follows:

$$r_i = \boldsymbol{p}_i \prod_{i'=1}^{i-1} r_{i'}^{-M(\boldsymbol{p}_i, \eta_{i'})}. \tag{23}$$

By construction, $M(r_i, \eta_j) = 0$ for $j < i$ and $M(r_i, \eta_i) = 1$. Moreover, $M(r_i, \eta_j) = 0$ for $j > i$ since

$$M(\boldsymbol{p}_i, \eta_j) = M(r_1, \eta_j) = \cdots = M(r_{i-1}, \eta_j).$$

Define $s_i(y) = \prod_{j=1}^{\ell} r_i(y_j)$ for $i = 1, \ldots, \ell$ so that s_i is the product of r_i evaluated at each indeterminate, just as q_i is the product of $\boldsymbol{p}_i$ evaluated at each indeterminate. In fact, we have

$$s_i = q_i \prod_{i'=1}^{i-1} s_{i'}^{-M(\boldsymbol{p}_i, \eta_{i'})}.$$

Let Q and S be the following mappings:

$$(y_1, \ldots, y_\ell) \overset{Q}{\mapsto} (q_1, \ldots, q_\ell) \overset{S}{\mapsto} (s_1, \ldots, s_\ell).$$

Consider the Jacobian of $S \circ Q$, evaluated at the point $\eta = (\eta_1, \ldots, \eta_\ell)$. By construction

$$\frac{\partial s_i}{\partial y_j}(\eta) = \left(\prod_{j' \neq j} r_i(\eta_{j'}) \right) r_i'(\eta_j). \tag{24}$$

Now

$$\prod_{j' \neq j} r_i(\eta_{j'}) \neq 0 \iff i = j$$

since $r_i(\eta_i) = 0$ and $M(r_i, \eta_j) = 0$ for any $j \neq i$. Moreover, $r_i'(\eta_i) \neq 0$ since η_i is a simple root of r_i, so we conclude that $\frac{\partial s_i}{\partial y_j}(\eta) \neq 0 \iff i = j$. Thus, the Jacobian is a diagonal matrix with non-zero diagonal entries and is therefore invertible. The Proposition follows. □

A.3 General condition for applying root and pole information

Here we provide a more general version of Proposition 5. Observe that the process (23) is effectively Gaussian elimination on the rows of the matrix M, which starts out lower-triangular with 1's on the diagonal. Carried further this yields:

Theorem 17. *Let $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_\ell$ be univariate rational functions and let $q_i(y) \coloneqq \prod_{j=1}^{\ell} \boldsymbol{p}_i(y_j)$ for $i = 1, \ldots, \ell$. Let $\eta_1, \ldots, \eta_L$ be the points which are roots or poles of any $\boldsymbol{p}_i$. Then the mapping $(y_1, \ldots, y_\ell) \mapsto (q_1(y), \ldots, q_\ell(y))$ is locally identifiable if and only if the $\ell \times L$ matrix $M(\boldsymbol{p}, \eta)$ with (i, j)'th entry $M(\boldsymbol{p}_i, \eta_j)$, has rank ℓ over $\mathbb{Q}$.*

Proof. Only If: Let $v \in \mathbb{Z}^\ell$ be a linear dependence of the rows, $v \cdot M(\boldsymbol{p}, \eta) = 0$. Then $\prod_{i'=1}^{\ell} q_{i'}(y)^{v_{i'}}$ factors as $\prod_j \prod_{i'=1}^{\ell} \boldsymbol{p}_{i'}(y_j)^{v_{i'}}$. By construction $\prod_{i'=1}^{\ell} \boldsymbol{p}_{i'}(\eta_j)^{v_{i'}}$ is nonzero and finite for every $1 \leq j \leq L$. Furthermore $\prod_{i'=1}^{\ell} \boldsymbol{p}_{i'}(x)^{v_{i'}}$ is nonzero and finite for all $x \notin \{\eta_1, \ldots, \eta_L\}$ because for such x every $\boldsymbol{p}_{i'}(x)$ is nonzero and finite. Thus, $\prod_{i'=1}^{\ell} \boldsymbol{p}_{i'}^{v_{i'}}$ is a rational function without finite roots or poles, and therefore a nonzero constant. So $\prod_{i'=1}^{\ell} q_{i'}(y)^{v_{i'}}$ is a nonzero constant. Consequently, the parameterized variety $q(y) = (q_1(y), \ldots, q_\ell(y))$ has codimension at least 1 in $\mathbb{C}^\ell$.

If: Without loss of generality suppose the submatrix of $M(\boldsymbol{p}, \eta)$ in columns $1, \ldots, \ell$ is nonsingular. Let N be a matrix with integer entries such that $N \cdot M(\boldsymbol{p}, \eta) = (D \mid B)$ where D is a diagonal matrix with positive integer entries on the diagonal, and B is any $\ell \times (L - \ell)$ matrix; $\mid$ denotes concatenation. Define the rational functions

$$r_i = \prod_{i'=1}^{\ell} \boldsymbol{p}_{i'}^{N_{ii'}} \tag{25}$$

By construction, for $j \leq \ell$, $M(r_i, \eta_j) = D_{ij}$. Define $s_i(y) \coloneqq \prod_{j=1}^{\ell} r_i(y_j)$ for $i = 1, \ldots, \ell$ so that s_i is the product of r_i evaluated at each indeterminate, just as q_i is the product of $\boldsymbol{p}_i$ evaluated at each indeterminate. Then

$$s_i(y) = \prod_{i'=1}^{\ell} q_{i'}(y)^{N_{ii'}}$$

Unlike in the proof of Prop. 5, it is not sufficient to consider the Jacobian of the mapping

$$(y_1, \ldots, y_\ell) \mapsto (s_1, \ldots, s_\ell)$$

because this Jacobian is singular if any $D_{ii} > 1$. However, we show the mapping is dimension-preserving by examining its expansion in a small neighborhood of $(\eta_1, \ldots, \eta_\ell)$. Observe, as in (24), that

$$\frac{\partial^k s_i}{\partial^k y_j}(\eta) = \left(\prod_{j' \neq j} r_i(\eta_{j'}) \right) r_i^{(k)}(\eta_j). \tag{26}$$

with $r_i^{(k)}$ being the k'th derivative of r_i. More generally, if $\vec{k} = (k_1, \ldots, k_\ell)$, $|\vec{k}| = \sum k_j$, $s_i^{(\vec{k})}(y) = \frac{\partial^{k_1}}{\partial^{k_1} y_1} \cdots \frac{\partial^{k_\ell}}{\partial^{k_\ell} y_\ell} s_i(y)$, then

$$s_i^{(\vec{k})}(\eta) = \prod_j r_i^{(k_j)}(\eta_j).$$

Let $s = (s_1, \ldots, s_\ell)$. In a small neighborhood of η, s expands in terms of those nonzero partial derivatives $(\vec{k})$ for which $\vec{k}$ is minimal in the standard partial order on the nonnegative quadrant. For each s_i this minimizer is unique, $(0, \ldots, 0, D_{ii}, 0, \ldots, 0)$. Thus (applying (26)), for small $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_\ell)$, $s(\eta + \varepsilon)$ expands as

$$(s_1(\eta), \ldots, s_\ell(\eta)) + \left(\left(\prod_{j' \neq 1} r_1(\eta_{j'}) \right) r_1^{(D_{11})}(\eta_1) \varepsilon_1^{D_{11}}, \ldots, \left(\prod_{j' \neq \ell} r_\ell(\eta_{j'}) \right) r_\ell^{(D_{\ell\ell})}(\eta_\ell) \varepsilon_\ell^{D_{\ell\ell}} \right)$$

This mapping carries ε in a small open neighborhood of 0 in $\mathbb{C}^\ell$, onto an open neighborhood of $s(\eta)$. □

A.4 Derivation for Equation 8

Observe that $r_0 = 1$ and recall that we have set $r_1 = \gamma$. Note that $2(\gamma - \sigma d) = \alpha_1 + \alpha_0$ and $(\gamma - \sigma d)^2 - d^2 = \alpha_1 \alpha_0$.

$$\begin{aligned} r_n &= \frac{\alpha_1^n + \alpha_0^n}{2} + \sigma \frac{\alpha_1^n - \alpha_0^n}{2} \\ &= (\alpha_1 + \alpha_0) \left(\frac{\alpha_1^{n-1} + \alpha_0^{n-1}}{2} + \sigma \frac{\alpha_1^{n-1} - \alpha_0^{n-1}}{2} \right) - \alpha_0 \alpha_1 \left(\frac{\alpha_1^{n-2} + \alpha_0^{n-2}}{2} + \sigma \frac{\alpha_1^{n-2} - \alpha_0^{n-2}}{2} \right) \\ &= (\alpha_1 + \alpha_0) r_{n-1} - \alpha_0 \alpha_1 r_{n-2} = 2(\gamma - \sigma d) r_{n-1} - [(\gamma - \sigma d)^2 - d^2] r_{n-2} \end{aligned}$$

A.5 Proof of Proposition 6, Section 3.2

Proof. Fix n, and induct on k. $k = 0$ is immediate from the definitions. The proofs for r_n and p_n are essentially identical as they rely only on the three-term recurrences (which are the same) and on the initial conditions p_{-1} and p_0. We write out the argument for r_n: it amounts to showing that the expression for k equals that for $k+1$:

$$\begin{aligned} p_k r_{n-k-1} - (x^2 - y) p_{k-1} r_{n-k-2} &= p_k (2x r_{n-k-2} - (x^2 - y) r_{n-k-3}) - (x^2 - y) p_{k-1} r_{n-k-2} \\ &= r_{n-k-2} (2x p_k - (x^2 - y) p_{k-1}) - (x^2 - y) p_k r_{n-k-3} \\ &= p_{k+1} r_{n-k-2} - (x^2 - y) p_k r_{n-k-3}. \end{aligned}$$

□

A.6 Proof of Lemma 7 (Zeroes of r_n on $x = 0, x^2 = y$)

Proof. First, for $x = 0$ the recursion takes the form $r_n(0, y) = y r_{n-2}$, and so:

$$r_n(0, y) = \begin{cases} \gamma y^{\lfloor n/2 \rfloor} & \text{if } n \equiv 0 \mod 2 \\ y^{\lfloor n/2 \rfloor} & \text{if } n \equiv 1 \mod 2 \end{cases}$$

This implies that if $r_n(0, y) = 0$, then $y = 0$. (Notice that this also forces $n \geq 2$.)

Second, for $x^2 = y$: here the recursion takes the form $r_n(x, x^2) = 2x r_{n-1}(x, x^2)$, and therefore we have $r_n(x, x^2) = \gamma (2x)^{n-1}$. Therefore if $r_n(x, x^2) = 0$ for $n \geq 2$ then $x = 0$. □

A.7 Proof of Lemma 11, (Properties of s_n polynomials on the $x = \gamma/2$)

Proof. Clearly, these statements are true of s_1 and s_2. Now suppose $n > 2$. For Part 1, observe that the leading coefficient of γs_{n-1} is positive by the inductive hypothesis, and the same is true of $y s_{n-2} - (\gamma^2/4) s_{n-2}$. Thus the leading coefficient of s_n is positive. For Part 2, observe:

$$s_n(0) = \gamma s_{n-1}(0) - \frac{\gamma^2}{4} s_{n-2}(0) = \gamma (n-1) (\frac{\gamma}{2})^{n-2} - \frac{\gamma^2}{4} (n-2) (\frac{\gamma}{2})^{n-3} = n (\frac{\gamma}{2})^{n-1}$$

For Part 3, first suppose n is odd. Then s_{n-1}, s_{n-2} have the same degree and so s_n has the same degree as ys_{n-2}, which is $(n-1)/2$. If n is even, then s_{n-1} has degree $(n-2)/2$ and s_{n-2} has degree $(n-4)/2$. Since the leading coefficients of γs_{n-1} and $(y-\gamma^2/4)s_{n-2}$ have the same sign by the inductive hypothesis, the degree of s_n is $(n-2)/2$. □

A.8 Proof of Lemma 12 (Roots of s_n are simple)

Proof. It is easy to check that $\beta_{3,1} = -\frac{3}{4}, \beta_{4,1} = -\frac{1}{4}$, so $\beta_{3,1} < \beta_{4,1}$, and they are both contained in the interval $(-\infty, 0)$.

We proceed by induction for all $n > 1$, treating (a), (b) separately.

(a) Observe that for $i \in [n-1]$,

$$s_{2n+1}(\beta_{2n,i}) = -\left(\frac{\gamma^2}{4} - \beta_{2n,i}\right) s_{2n-1}(\beta_{2n,i}).$$

By the inductive hypothesis, $\beta_{2n,i} < 0$, so $\frac{\gamma^2}{4} - \beta_{2n,i}$ is positive, and for convenience we will denote it by c_i. Now note as we range over all i, the sign of $s_{2n-1}(\beta_{2n,i})$ changes every time we increment i because s_{2n-1} interlaces s_{2n} by the inductive hypothesis. By the Intermediate Value Theorem, we have found $n-2$ roots in the intervals $(\beta_{2n,i}, \beta_{2n,i+1})$ for $i = 1, \dots, n-2$. We have two more roots to account for. Note

$$\operatorname{sign}(s_{2n+1}(\beta_{2n,1})) = -\operatorname{sign}(s_{2n-1}(\beta_{2n,1})) = \operatorname{sign}\left(\lim_{v\to-\infty} s_{2n-1}(v)\right) = -\operatorname{sign}\left(\lim_{v\to-\infty} s_{2n+1}(v)\right).$$

The first equality holds by the recurrence relation; the second equality holds because $\beta_{2n-1,1} < \beta_{2n,1}$, and the third equality holds because the degrees of s_{2n-1}, s_{2n+1} are different by 1. Thus, s_{2n+1} has an odd number of roots, and therefore one root, in the interval $(-\infty, \beta_{2n,1})$. Finally, observe that $s_{2n-1}(\beta_{2n,n-1}) > 0$ because $\beta_{2n-1,n-1} < \beta_{2n,n-1}$ and s_{2n-1} has positive leading coefficient by the previous lemma. Thus, $s_{2n+1}(\beta_{2n,n-1}) < 0$. Since $s_{2n+1}(0) > 0$, s_{2n+1} has a root in the interval $(\beta_{2n,n-1}, 0)$. Thus we've accounted for all n roots of s_{2n+1}, and shown that they are all negative and interlace the roots of s_{2n}.

(b) We now show that the roots of s_{2n+2} interlace those of s_{2n+1} and 0. First, observe that $s_{2n+2}(\beta_{2n+1,i}) = -c_i s_{2n}(\beta_{2n+1,i})$ for $i = 1, \dots, n$, where we have made the obvious change of definition for $c_i > 0$. Since $s_{2n}(\beta_{2n+1,i})$ changes sign every time we increment i, by the Intermediate Value Theorem, s_{2n+2} has at least one root each in $(\beta_{2n+1,i}, \beta_{2n+1,i+1})$ for $i = 1, \dots, n-1$. Finally, we can see that $s_{2n}(\beta_{2n+1,n}) > 0$ because $\beta_{2n,n-1} < \beta_{2n+1,n}$ and s_{2n} has positive leading coefficient. Thus, $s_{2n+2}(\beta_{2n+1,n}) < 0$, so s_{2n+2} has a root in the interval $(\beta_{2n+1,n}, 0)$. We have now accounted for all n roots of s_{2n+2}. □

A.9 Proof of Lemma 14 (Degrees of atomic polynomials)

Proof. From (11), $\deg s_n = \sum_{d|n} f(d)$. Next perform Möbius inversion in the division lattice to obtain an expression for $f(n)$ in terms of $F(d) := \deg s_d = \lfloor (d-1)/2 \rfloor$: that is, for μ the Möbius function of the division lattice, $f(n) = \sum_{d|n} F(d)\mu(n/d)$. Letting the prime factorization of n be $n = q_1^{\beta_1} \cdots q_k^{\beta_k}$ with $q_i < q_{i+1}$, this expression simplifies to $f(n) = \sum_{S\subseteq[k]} (-1)^{|S|} F(n/q^S)$ where $q^S := \prod_{i\in S} q_i$. Now consider three cases.

First, suppose n is odd. Then $\lfloor (n-1)/2 \rfloor = (n-1)/2$, and n/q^S is odd for any S. Observe

$$f(n) = \frac{1}{2} \sum_{S\subseteq[k]} (-1)^{|S|} \left(\frac{n}{q^S} - 1\right) = \frac{n}{2} \sum_{S\subseteq[k]} (-1)^{|S|} \frac{1}{q^S} = \frac{n}{2} \prod_{i=1}^{k} \left(1 - \frac{1}{q_i}\right).$$

The second equality follows because S has as many even-sized as odd-sized subsets.

Second, suppose that $4 \mid n$. Now n/q^S is even for any S because q^S contains at most one factor of 2. For even n, $\lfloor (n-1)/2 \rfloor = (n-2)/2$. The argument now follows the pattern for n odd.

Third, suppose that $n = 2m$, $m > 1$ odd. Now $q_1 = 2, \beta_1 = 1$. So for $S \subset [k]$ if $1 \notin S$ then $F(n/q^S) = \frac{n}{q^S} - 2$, and if $1 \in S$ then $F(n/q^S) = \frac{n}{q^S} - 1$.

$$f(n) = \frac{1}{2}\sum_{1\notin S}(-1)^{|S|}\left(\frac{n}{q^S}-2\right) + \frac{1}{2}\sum_{1\in S}(-1)^{|S|}\left(\frac{n}{q^S}-1\right)$$
$$= \frac{n}{2}\sum_{1\notin S}(-1)^{|S|}\frac{1}{q^S} - \frac{n}{2}\sum_{1\notin S}(-1)^{|S|}\frac{1}{2q^S} = \frac{n}{4}\prod_{i=2}^{k}\left(1-\frac{1}{q_i}\right) = \frac{n}{2}\prod_{i=1}^{k}\left(1-\frac{1}{q_i}\right).$$

□

A.10 Proof of Lemma 15 (Formula for Jacobian)

Proof. Pick any point on the line $v = (\gamma/2, v_0)$. Notice that we can rewrite r_i and p_{i-1} as follows:

$$p_{i-1} = 2xp_{i-2} - (x^2-y)p_{i-3}$$
$$r_i = \gamma p_{i-2} - (x^2-y)p_{i-3}$$

Writing the partials of both equations we see:

$$\frac{\partial r_i}{\partial y} = \gamma\frac{\partial p_{i-2}}{\partial y} - (x^2-y)\frac{\partial p_{i-3}}{\partial y} + p_{i-3}$$
$$\frac{\partial r_i}{\partial x} = \gamma\frac{\partial p_{i-2}}{\partial x} - (x^2-y)\frac{\partial p_{i-3}}{\partial x} - 2xp_{i-3}$$
$$\frac{\partial p_{i-1}}{\partial y} = 2x\frac{\partial p_{i-2}}{\partial y} - (x^2-y)\frac{\partial p_{i-3}}{\partial y} + p_{i-3}$$
$$\frac{\partial p_{i-1}}{\partial x} = 2x\frac{\partial p_{i-2}}{\partial x} - (x^2-y)\frac{\partial p_{i-3}}{\partial x} - 2xp_{i-3} + 2p_{i-2}$$

Since we are only interested in the solutions on the line $x = \frac{\gamma}{2}$, we can now rewrite the following partials:

$$\frac{\partial p_{i-1}}{\partial y}(v) = \frac{\partial r_i}{\partial y}(v)$$
$$\frac{\partial p_{i-1}}{\partial x}(v) = \frac{\partial r_i}{\partial x}(v) + 2s_i(v_0)$$

Finally we see that:

$$\frac{\partial p_{i-1}}{\partial x}(v)\frac{\partial r_i}{\partial y}(v) - \frac{\partial p_{i-1}}{\partial y}(v)\frac{\partial r_i}{\partial x}(v) = \left(\frac{\partial r_i}{\partial x}(v) + 2s_i(v_0)\right)\frac{\partial r_i}{\partial y}(v) - \frac{\partial r_i}{\partial y}(v)\frac{\partial r_i}{\partial x}(v)$$
$$= 2s_i(v_0)\frac{\partial r_i}{\partial y}(v) = 2s_i(v_0)\frac{\partial s_{i+1}}{\partial y}(v_0)$$

We know on the line $s_{i+1} = r_i = p_{i-1}$ for all i, so plugging this back into Equation 12, we get the following expression as desired:

$$\det(J_i)(v) = -2s_{i+2}(v_0)s_i(v_0)\frac{\partial s_{i+1}}{\partial y}(v_0) + s_{i+1}(v_0)(F_i(v) + G_i(v))$$

□

A.11 Proof of Theorem 16, Section 3.7

Proof. From Theorem 10 we know $s_{2n+1} \mid s_{4n+2}$, let c_n be an atomic root of s_{2n+1} and let $C = \{c_n\}_{n=1}^{\ell}$. For ease of notation, let $R_n = \{j \mid s_{2n}(c_j) = 0\}$ and $|R_n| = \alpha_n$. Let us make the following observations about this set:

Observation 18. *For all $j \in R_n$, $j \le n$. Moreover, $s_{4n+2}(c_j) = 0$ if and only if $j \in R_n$*

Proof. Suppose $j > n$, we know c_j is an atomic root of s_{2j+1} and by definition of an atomic root $s_{2n+1}(c_j) \neq 0$, thus if $j \in R_n$ then $j \leq n$.

Clearly if $j \in R_n$ then $s_{4n+2}(c_j) = 0$. If $s_{4n+2}(c_j) = 0$, since c_j is an atomic root of s_{2j+1}, then $\gcd(s_{4n+2}, s_{2j+1}) = s_{2j+1}$. This implies $2j+1 \mid 4n+2$ and thus $2j+1 \mid 2n+1$ and it follows $s_{2n+1}(c_j) = 0$ from Theorem 10 so $j \in R_n$. □

We know that each root of s_{2n+1} is simple for all n, thus $(\partial s_{2n+1}/\partial y)(c_i) \neq 0$. Furthermore, since $s_{2n+1}(c_n) = 0$ then we know that $s_{2n}(c_n) \neq 0$ and $s_{2n+2}(c_n) \neq 0$. Lastly, notice that from the above observation, $i \in R_n$ if and only if $s_{4n+2}(c_i) = 0$. Together with Theorem 10 this implies that there exists some $\delta > 0$ such that for all n, we have that s_{2n}, s_{2n+2}, $\partial s_{2n+1}/\partial y$, and s_{2i+1} and s_{4i+2} for all $i \in \{1, \ldots, \ell\} \setminus R_n$, are all bounded away from zero in the interval $I_n = [c_n - \delta, c_n + \delta]$, by some constant A.

Clearly $T = \cup_{n=1}^{\ell} I_n$ is closed and bounded, then so is the set $\{\gamma/2\} \times T$ and thus each of the functions in the following set attain a maximum over $\{\gamma/2\} \times T$:

$$\bigcup_{n=1}^{\ell} \left\{ |r_{2n}|, |r_{2n+1}|, |r_{2n-1}|, |\frac{\partial r_{2n}}{\partial x}|, |\frac{\partial r_{2n}}{\partial y}|, \right. \\ \left. |\frac{\partial r_{4n+1}}{\partial x}|, |\frac{\partial r_{4n+1}}{\partial y}|, |F_{2n}|, |G_{2n}| \right\}$$

Define M to be the maximum over the maximums of these functions and 1.

We now pick some small $\epsilon > 0$ to be specified later. Define the set $D_j = \{n \mid j \in R_n\}$. We will pick our points as follows, for all $1 \leq i \leq \ell$, we pick $d_i \in I_i$ such that $0 < s_{2k+1}(d_i) < \epsilon_i$ and $0 < s_{4k+2}(d_i) < \epsilon_i$ for all $k \in D_i$. If $1 \leq i < \ell$, we define ϵ_{i+1} as follows:

$$\epsilon_{i+1} = \min \left(\bigcup_{k \in D_i} \{|s_{2k+1}(d_i)|, |s_{4k+2}(d_i)|\} \right)$$

Notice that this process results in a set of points $\{d_i\}$ where for all n and $k \in R_n$ if $k \neq n$ then $s_{2n+1}(d_n) < \epsilon_n < s_{2n+1}(d_k)$ and $s_{4n+2}(d_n) < \epsilon_n < s_{4n+2}(d_k)$. Therefore for all $k \in R_n$:

$$\left| \frac{s_{2n+1}(d_n)}{s_{2n+1}(d_k)} \right| \leq 1 \qquad \left| \frac{s_{4n+2}(d_n)}{s_{4n+2}(d_k)} \right| \leq 1 \tag{27}$$

We now evaluate the Jacobian at the point $(\gamma/2, d_1, \ldots, \gamma/2, d_\ell)$. We scale the rows corresponding to q_{2n} and q_{4n+1} by the following two non-zero values respectively:

$$s_{2n+1}(d_n) \prod_{k \in R_n} \frac{1}{s_{2n+1}(d_k)} \qquad s_{4n+2}(d_n) \prod_{k \in R_n} \frac{1}{s_{4n+2}(d_k)}$$

We call the resulting matrix B, and notice that B is non-singular if and only if the Jacobian evaluated at this point is non-singular. For ease of notation, we will refer to the i, jth entry of the matrix B as $b_{i,j}$ and we will split the matrix B into 2×2 blocks.

$$B = \begin{pmatrix} N_{1,1} & \ldots & N_{1,\ell} \\ \vdots & & \vdots \\ N_{\ell,1} & \ldots & N_{\ell,\ell} \end{pmatrix} \tag{28}$$

Notice that each 2×2 block has a similar structure, take any $n, m \in [\ell]$ and we can explicitly write the matrix corresponding to $N_{n,m}$.

$$N_{n,m} = \begin{pmatrix} \frac{s_{2n+1}(d_n)}{s_{2n+1}(d_m)} \prod_{k \notin R_n} s_{2n+1}(d_k) & 0 \\ 0 & \frac{s_{4n+2}(d_n)}{s_{4n+2}(d_m)} \prod_{k \notin R_n} s_{4n+2}(d_k) \end{pmatrix} \begin{pmatrix} \frac{\partial r_{2n}}{\partial x}(\frac{\gamma}{2}, d_m) & \frac{\partial r_{2n}}{\partial y}(\frac{\gamma}{2}, d_m) \\ \frac{\partial r_{4n+1}}{\partial x}(\frac{\gamma}{2}, d_m) & \frac{\partial r_{4n+1}}{\partial y}(\frac{\gamma}{2}, d_m) \end{pmatrix}$$

Lemma 19. *Suppose $b_{i,j} \in N_{n,m}$. If $m \in R_n$ then $|b_{i,j}| \leq M^\ell$, otherwise $|b_{i,j}| \leq \epsilon M^{\ell-1}$*

Proof. First we will make the following observation, assume that i and j are odd, since $M \geq 1$ we have that:

$$\left|\frac{\partial r_{2n}}{\partial x}(d_m)\right| \prod_{k \notin R_n} |s_{2n+1}(d_k)| < M^\ell \tag{29}$$

If i is even we replace $2n$ with $4n+1$, and if j is even we replace x with y. Notice that the same argument works in all of those cases. If $m \in R_n$, then we know that Equation 27 implies that both $|s_{2n+1}(d_n)/s_{2n+1}(d_m)| \leq 1$ and $|s_{4n+2}(d_n)/s_{4n+2}(d_m)| \leq 1$. Together with Equation 29 this implies that $|b_{i,j}| < M^\ell$ as desired.

Suppose that $m \notin R_n$, we will do the following analysis assuming both i and j are odd, but an identical argument works for either i and j even.

$$\frac{\partial r_{2n}}{\partial x}(\gamma/2, d_m) \cdot \frac{s_{2n+1}(d_n)}{s_{2n+1}(d_m)} \prod_{k \notin R_n} s_{2n+1}(d_k) = \frac{\partial r_{2n}}{\partial x}(\gamma/2, d_m) \cdot s_{2n+1}(d_n) \prod_{k \notin R_n \cup \{m\}} s_{2n+1}(d_k)$$

Notice $|s_{2n+1}(d_n)| < \epsilon_n < \epsilon$ and all other terms are bounded above in magnitude by M, since there are at most $\ell - 1$ of those terms and $M > 1$ we have that $|b_{i,j}| < \epsilon M^{\ell-1}$ as desired. □

The intuition for the rest of the proof is as follows. We know that for a block lower-triangular matrix, the determinant is the product of its diagonal blocks. This is because any permutation π which selects an element of the lower-left block must also select an element from the zero block and so the term associated to π contributes nothing to the determinant.

Similar to this, we will show that the product of the determinant of the diagonal blocks, $N_{n,n}$, contains a term which is not dependent on ϵ. In the following lemma we will show that all π which pick elements of off-diagonal blocks scale with ϵ. This will imply for sufficiently small ϵ this matrix must be non-singular.

Let $H \subset S_{2\ell}$ be the subgroup generated by the ℓ transpositions $(2n-1, 2n)$ for $n \in [\ell]$.

Lemma 20. *For all $\pi \in S_{2\ell} \setminus H$:*

$$\left|\operatorname{sign}(\pi) \prod_{i=1}^{\ell} b_{i,\pi(i)}\right| < \epsilon M^{\ell^2 - 1}$$

Proof. Pick any $\pi \in S_{2\ell}$, notice that if for all $i \in \{1, \ldots, \ell\}$, $\pi(2i) \leq 2i$ and $\pi(2i-1) \leq 2i$ then $\pi \in H$. Thus since we pick $\pi \in S_{2\ell} \setminus H$, then for some i either $\pi(2i) > 2i$ or $\pi(2i-1) > 2i$, in both cases we have that for some j, $b_{j,\pi(j)}$ lies in $N_{n,m}$ for $m > n$.

From Observation 18, we noted that if $m > n$, then $m \notin R_n$ and thus from Lemma 19 $|b_{j,\pi(j)}| < \epsilon M^{\ell-1}$ and for all other $k \in \{1, \ldots, \ell\} \setminus \{j\}$, $|b_{k,\pi(k)}| < M^\ell$. From this, the inequality follows trivially. □

Now we turn our attention to the portion of the determinant contributed by all the permutations in H.

$$\sum_{\pi \in H} \operatorname{sign}(\pi) \prod_{i=1}^{\ell} b_{i,\pi(i)} = \prod_{n=1}^{\ell} \det(N_{n,n}) \tag{30}$$

$$= \prod_{n=1}^{\ell} \det(J_{2n}(d_n)) \left(\prod_{k \notin R_n} s_{2n+1}(d_k) \cdot s_{4n+2}(d_k) \right) \tag{31}$$

Notice that using the definitions for M and A, we can bound the magnitude of the following product from above and below.

$$A^{2\ell^2 - 2(\alpha_1 + \cdots + \alpha_\ell)} \leq \left| \prod_{n=1}^{\ell} \prod_{k \notin R_n} s_{2n+1}(d_k) \cdot s_{4n+2}(d_k) \right| \leq M^{2\ell^2} \tag{32}$$

Lemma 21.

$$\left|\prod_{n=1}^{\ell} \det(J_{2n}(d_n))\right| \geq 2^\ell A^{3\ell} - \epsilon 2^\ell (2^\ell - 1) M^{3\ell-2}$$

Proof. From Lemma 15:

$$\prod_{n=1}^{\ell} \det(J_{2n}(d_n)) = \prod_{n=1}^{\ell} (-2 s_{2n+2}(d_n) s_{2n}(d_n) \frac{\partial s_{2n+1}}{\partial y}(d_n) + s_{2n+1}(d_n)(F_{2n}(\gamma/2, d_n) + G_{2n}(\gamma/2, d_n)))$$

We will first bound the magnitude of the term of this product that results from picking the left term of each factor. From the definition of A we get the following bound.

$$\left|\prod_{n=1}^{\ell} -2 s_{2n+2}(d_n) s_{2n}(d_n) \frac{\partial s_{2n+1}}{\partial y}(d_n)\right| \geq 2^\ell A^{3\ell}$$

All of the rest of the terms must include $s_{2n+1}(d_n)$ for some n, and we know that $|s_{2n+1}(d_n)| < \epsilon$ for all n. Using this fact and the definition of M, we observe the following inequalities.

$$|s_{2n+1}(d_n)(F_{2n}(\gamma/2, d_n) + G_{2n}(\gamma/2, d_n)| < \epsilon(2M)$$
$$|-2 s_{2n+2}(d_n) s_{2n}(d_n) \frac{\partial s_{2n+1}}{\partial y}(d_n)| < 2M^3$$

Clearly $2M^3 > \epsilon(2M)$ so each of the $2^\ell - 1$ other terms in the product are bounded above in magnitude by $\epsilon 2^\ell M^{3\ell-2}$. Thus the magnitude of the product of the determinants is bounded from below by $2^\ell A^{3\ell} - \epsilon 2^\ell (2^\ell - 1) M^{3\ell-2}$ as desired. □

Using Lemma 20, Lemma 21, and Equation 32 we bound the magnitude of the determinant of B from below.

$$\begin{aligned}
|\det(B)| &\geq \left|\sum_{\pi \in H} \operatorname{sign}(\pi) \prod_{i=1}^{\ell} b_{i,\pi(i)}\right| - \left|\sum_{\pi \in S_{2\ell} \setminus H} \operatorname{sign}(\pi) \prod_{i=1}^{\ell} b_{i,\pi(i)}\right| \\
&\geq \left(\prod_{k \notin R_n} s_{2n+1}(d_k) \cdot s_{4n+2}(d_k)\right) \left(2^\ell A^{3\ell} - \epsilon 2^\ell (2^\ell - 1) M^{3\ell-2}\right) - \epsilon (2\ell)! M^{\ell^2 - 1} \\
&\geq 2^\ell A^{2\ell^2 + 3\ell - 2(\alpha_1 + \cdots + \alpha_\ell)} - \epsilon \left(2^\ell (2^\ell - 1) M^{2\ell^2 + 3\ell - 2} (2\ell)! M^{\ell^2 - 1}\right)
\end{aligned}$$

Notice that for ϵ sufficiently small, we have that $|\det(B)| > 0$ and therefore the Jacobian of our desired map is non-singular. This implies that our desired map is locally identifiable, concluding the proof. □