This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3146630

Hadamard Extensions and the Identification of Mixtures of Product Distributions

Spencer L. Gordon and Leonard J. Schulman

***Abstract*—The Hadamard Extension $\mathbb{H}(\mathbf{m})$ of an $n \times k$ matrix $\mathbf{m}$ is the collection of all Hadamard products of subsets of its rows. This construction is essential for source identification (parameter estimation) of a mixture of k product distributions over n binary random variables. A necessary requirement for such identification is that $\mathbb{H}(\mathbf{m})$ have full column rank; conversely, identification is possible if apart from each row there exist two disjoint sets of rows of $\mathbf{m}$, each of whose extension has full column rank. It is necessary therefore to understand when $\mathbb{H}(\mathbf{m})$ has full column rank; we provide two results in this direction. The first is that if $\mathbb{H}(\mathbf{m})$ has full column rank then there exists a set of at most $k-1$ rows of $\mathbf{m}$, whose extension already has full column rank. The second is a Hall-type condition on the values in the rows of $\mathbf{m}$, that suffices to ensure full column rank of $\mathbb{H}(\mathbf{m})$.**

***Index Terms*—Machine Learning Algorithms, Mixture Models, Parameter Estimation.**

I. Introduction

The Hadamard product for row vectors $u = (u_1, \ldots, u_k)$, $v = (v_1, \ldots, v_k)$ is the mapping $\odot : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k$ given by

$$u \odot v := (u_1 v_1, \ldots, u_k v_k)$$

The identity for this product is the all-ones vector $\mathbb{1}$. We associate with vector v the linear operator $v_\odot = \mathrm{diag}(v)$, a $k \times k$ diagonal matrix, so that

$$u \cdot v_\odot = v \odot u.$$

Throughout this paper $\mathbf{m}$ is a real matrix with row set $[n] := \{1, \ldots, n\}$ and column set $[k]$; write $\mathbf{m}_i$ for a row and $\mathbf{m}^j$ for a column.

As a matter of notation, for a matrix Q and nonempty sets R of rows and C of columns, let $Q|_R^C$ be the restriction of Q to those rows and columns (with either index omitted if all rows or columns are retained).

Definition 1. The Hadamard Extension of $\mathbf{m}$, written $\mathbb{H}(\mathbf{m})$, is the $2^n \times k$ matrix with rows $\mathbf{m}_S$ for all $S \subseteq [n]$, where, for $S = \{i_1, \ldots, i_\ell\}$, $\mathbf{m}_S = \mathbf{m}_{i_1} \odot \cdots \odot \mathbf{m}_{i_\ell}$; equivalently $\mathbf{m}_S^j = \prod_{i \in S} \mathbf{m}_i^j$. (In particular $\mathbf{m}_\emptyset = \mathbb{1}$.)

This construction originated recently in learning theory [3], [8] where it is arises naturally and unavoidably when we wish to perform source identification (i.e., parameter estimation) given data from a mixture (convex combination) of k product distributions on n binary random variables. We explain the connection further in Section II. Motivated by this application, we are interested in the following two questions:

(1) If $\mathbb{H}(\mathbf{m})$ has full column rank, must there exist a subset R of the rows, of bounded size, such that $\mathbb{H}(\mathbf{m}|_R)$ has full column rank?

(2) In each row of $\mathbf{m}$, assign distinct colors to the distinct real values. Is there a condition on the coloring that ensures $\mathbb{H}(\mathbf{m})$ has full column rank?

In answer to the first question we show:

Theorem 2. *If $\mathbb{H}(\mathbf{m})$ has full column rank then there is a set R of no more than $k-1$ of the rows of $\mathbf{m}$, such that $\mathbb{H}(\mathbf{m}|_R)$ has full column rank.*

Considering the more combinatorial second question, observe that if $\mathbf{m}$ possesses two identical columns then the same is true of $\mathbb{H}(\mathbf{m})$, and so the latter cannot have full column rank. Extending this further, suppose there are three columns C in which only one row r has more than one color. Then $\mathrm{rowspace}\,\mathbb{H}(\mathbf{m}|^C)$ is spanned by $\mathbb{1}|^C$ and $r|^C$, so again $\mathbb{H}(\mathbf{m})$ cannot have full column rank. Motivated by these necessary conditions, set:

Definition 3. For a matrix Q let $\mathrm{NAE}(Q)$ be the set of nonconstant rows of Q (NAE="not all equal"); let $\varepsilon(Q|^C) = |\mathrm{NAE}(Q|^C)| - |C|$; and let $\bar{\varepsilon}(Q) = \min_{C \neq \emptyset} \varepsilon(Q|^C)$. If $\bar{\varepsilon}(Q) \geq -1$ we say Q satisfies the NAE condition.

In answer to the second question we have the following:

Theorem 4. *If $\mathbf{m}$ satisfies the NAE condition then*

(a) There is a restriction of $\mathbf{m}$ to some $k-1$ rows R such that $\bar{\varepsilon}(\mathbf{m}|_R) = -1$.

(b) $\mathbb{H}(\mathbf{m})$ is full column rank.

(As a consequence also $\mathbb{H}(\mathbf{m}|_R)$ is full column rank.)

Apparently the only well-known example of the NAE condition is when $\mathbf{m}$ contains $k-1$ rows which are identical and whose entries are all distinct. Then the vectors $\mathbf{m}_\emptyset, \mathbf{m}_{\{1\}}, \mathbf{m}_{\{1,2\}}, \ldots, \mathbf{m}_{\{1,\ldots,k-1\}}$ form a nonsingular Vandermonde matrix. This example shows that the bound of $k-1$ in (a) is best possible.

For another example in which the NAE condition ensures that $\mathrm{rank}\,\mathbb{H}(\mathbf{m}) = k$, take the $(k-1)$-row matrix with $\mathbf{m}_i^j = 1$ for $i \leq j$ and $\mathbf{m}_i^j = 1/2$ for $i > j$. Here the NAE condition is only minimally satisfied, in that for every $\ell \leq k$ there are ℓ columns C s.t. $\varepsilon(\mathbf{m}|^C) = -1$.

For $k > 3$ the NAE condition is no longer necessary in order that $\mathbb{H}(\mathbf{m})$ have full column rank. E.g., for $k = 2^\ell$, the $\ell \times k$ "Hamming matrix" $\mathbf{m}_i^j = (-1)^{j_i}$ where j is an ℓ-bit string $j = (j_1, \ldots, j_\ell)$, forms $\mathbb{H}(\mathbf{m}) =$ the Fourier transform for the

Research supported in part by NSF grant CCF-1909972. The authors are with the Division of Engineering and Applied Science, California Institute of Technology, Pasadena CA 91125 USA (emails: slgordon, schulman@caltech.edu). A draft of this article was posted at ArXiv:2101.11688.

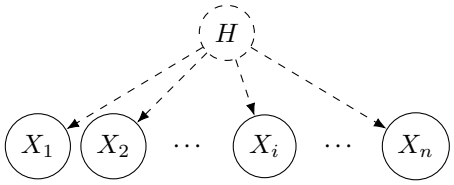
This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3146630

group $(\mathbb{Z}/2)^\ell$ (often called a Walsh or Hadamard transform), which is invertible.

Furthermore, for $k \leq 2^\ell$, almost all (in the sense of Lebesgue measure) $\ell \times k$ matrices $\mathbf{m}$ form a full-column-rank $\mathbb{H}(\mathbf{m})$. (For $k = 2^\ell$ this is because $\det \mathbb{H}(\mathbf{m})$ is a polynomial in the entries of $\mathbf{m}$, and the Walsh example shows that this polynomial is nonzero. For $k < 2^\ell$, consideration of the same $2^\ell \times 2^\ell$ Walsh transform implies that there are some k rows of $\mathbb{H}(\mathbf{m})$ such that the determinant of the minor they form is a nonzero polynomial.) Despite this observation, the Vandermonde case, in which $k-1$ rows are required, is very typical, as it is what arises in $\mathbb{H}(\mathbf{m})$ for a mixture model of observables X_i that are iid conditional on a hidden variable. Another class of examples that is far from Lebesgue-typical, and furthermore also far from being "separated" (see next section), is this. There are two possible coins, with biases $p_1 \neq p_2$. A hidden variable H is sampled in $\{0, \ldots, k-1\}$, and then the process is that you observe the result of H independent tosses of coin 1, followed by $k-1-H$ independent tosses of coin 2. The NAE condition implies that here $\mathbb{H}(\mathbf{m})$ has full column rank. As a consequence (applying [8]) the following model is identifiable: a hidden H is sampled (from unknown prior) in $\{0, \ldots, k-1\}$, and then you observe the result of $2H$ independent tosses of coin 1 followed by $2k-1-2H$ independent tosses of coin 2. A similar class of examples (sometimes identifiable but in general not) are the "subcube mixtures" studied in [3], where all coin biases must be one of $\{0, 1/2, 1\}$.

II. Motivation

Consider *observable* random variables $X_1, \ldots, X_n$ that are statistically independent conditional on H, a *hidden* or *latent* random variable H supported on $\{1, \ldots, k\}$. (See causal diagram.)



The most fundamental case is that the X_i are binary. Then we denote $\mathbf{m}_i^j = \Pr(X_i = 1 | H = j)$. The model parameters are $\mathbf{m}$ along with a probability distribution (the *mixture* distribution) $\pi = (\pi_1, \ldots, \pi_k)$ on H.

The study of finite mixture models was pioneered in the late 1800s in [13], [14]. The problem of learning such distributions has drawn a great deal of attention. For surveys see, e.g., [5], [17], [11], [12]. For some algorithmic papers on discrete-valued X_i, see [9], [4], [7], [2], [6], [1], [15], [10], [3], [8]. The source identification (or parameter estimation) problem is that of computing $(\mathbf{m}, \pi)$ from the joint statistics of the X_i. Put another way, the problem is to invert the multilinear moment map

$$\begin{aligned} &\mu : (\mathbf{m}, \pi) \to \mathbb{R}^{2^{[n]}} \\ &\mu(\mathbf{m}, \pi)_S = \Pr(X_S = 1) \quad \text{where } S \subseteq [n],\ X_S = \prod\nolimits_{i \in S} X_i \\ &\qquad\qquad = \mathbf{m}_S \cdot \pi^\top \end{aligned}$$

Since $\mathbf{m}_S^j = \Pr(X_S = 1 | H = j)$, this shows the essential role of $\mathbb{H}(\mathbf{m})$ in the mixture model.

Connection to $\operatorname{rank} \mathbb{H}(\mathbf{m})$

In general μ is not injective (even allowing for permutation among the values of π and columns of $\mathbf{m}$). For instance it is clearly not injective if $\mathbf{m}$ has two identical columns (unless π places no weight on those). More generally, and assuming all $\pi_j > 0$, it cannot be injective unless $\mathbb{H}(\mathbf{m})$ has full column rank. (Suppose $\alpha \in \mathbb{R}^k$ is nonzero s.t. $\mathbb{H}(\mathbf{m})\alpha = 0$. Since $(\mathbb{H}(\mathbf{m})\alpha)|_{\{\emptyset\}} = 0$, $\sum_j \alpha_j = 0$. So for sufficiently small $\delta > 0$, $\pi + \delta\alpha$ is a mixture distribution, distinct from π, with identical statistics.)

One sufficient condition for injectivity, due to [16], is that there be $2k-1$ "separated" observables X_i. X_i is separated if all $\mathbf{m}_i^j$ are distinct, or in our terminology, if no color recurs in $\mathbf{m}_i$. (Further it is shown in [8], Theorem 1, that one can lower bound the distance between $\mu(\mathbf{m}, \pi)$ and any $\mu(\mathbf{m}', \pi')$ in terms of $\zeta = \min_i \min_{j \neq j'} |\mathbf{m}_i^j - \mathbf{m}_i^{j'}|$ and the distance between $(\mathbf{m}, \pi)$ and $(\mathbf{m}', \pi')$.) There are examples with $X_1, \ldots, X_{2k-1}$ where the mapping is injective but is no longer so if any single X_i is omitted [15].

A weaker and still sufficient condition for injectivity of μ, due to [8], is that for every $i \in [n]$ there exist two disjoint sets $A, B \subseteq [n] - \{i\}$ such that $\mathbb{H}(\mathbf{m}|_A)$ and $\mathbb{H}(\mathbf{m}|_B)$ have full column rank. (It is not known whether two disjoint such A, B are strictly necessary.)

Observable X_i *with larger finite range*

If each X_i can take on one of say L values, $\mathbf{m}$ can be considered as a nonnegative $n \times k \times L$ real array, with $\mathbf{m}_{i,\ell}^j = \Pr(X_i = \ell | H = j)$, $\sum_{\ell=1}^L \mathbf{m}_{i,\ell}^j = 1$; the multivariate moments are indexed not by sets S but by mappings $S : [n] \to [L]$, with $\mathbf{m}_S = \mathbf{m}_{S(1)} \odot \cdots \odot \mathbf{m}_{S(n)}$ and

$$\begin{aligned} &\mu : (\mathbf{m}, \pi) \to \mathbb{R}^{[L]^{[n]}} \\ &\mu(\mathbf{m}, \pi)_S = \Pr(X_S = 1) \quad \text{where } X_S = \prod\nolimits_{i=1}^n \delta_{X_i, S(i)} \\ &\qquad\qquad \text{(Kronecker delta)} \\ &\qquad\qquad = \mathbf{m}_S \cdot \pi^\top \end{aligned}$$

For any given k, if L is sufficiently large and $\mathbf{m}$ satisfies a certain nonsingularity condition, the mixture learning problem becomes easier; this insight is due to [1]. It will be interesting to explore what conditions exactly $\mathbf{m}$ must satisfy for identifiability (for positive π), for arbitrary L. But in this paper we study only the most extreme, and hardest for identification, case $L = 2$.

III. Some Theory for Hadamard Products, and a Proof of Theorem 2

For $v \in \mathbb{R}^k$ and U a subspace, extend the definition of $v_\odot$ to

$$v_\odot(U) = \{u \cdot v_\odot : u \in U\}$$

and introduce the notation

$$v_{\bar{\odot}}(U) = \operatorname{span}(U \cup v_\odot(U)).$$

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3146630

We want to understand which subspaces U are invariant under $v_{\bar{\odot}}$. Let v have distinct values $\lambda_1 > \ldots > \lambda_\ell$ for $\ell \le k$. Let the polynomials $p_{v,i}$ $(i = 1, \ldots, \ell)$ of degree $\ell - 1$ be the Lagrange interpolation polynomials for these values, so $p_{v,i}(\lambda_j) = \delta_{ij}$ (Kronecker delta). Let $B(v)$ denote the partition of $[k]$ into blocks $B(v)_{(i)} = \{j : v_j = \lambda_i\}$. Let $V_{(i)}$ be the space spanned by the elementary basis vectors in $B(v)_{(i)}$, and $P_{(i)}$ the projection onto $V_{(i)}$ w.r.t. the standard inner product. Since $v_\odot$ is diagonal with entries λ_i in $B(v)_{(i)}$, we have the matrix equation

$$p_{v,i}(v_\odot) = P_{(i)}, \tag{1}$$

where $p_{v,i}$ is interpreted as a matrix polynomial. The collection of all linear combinations of the matrices $P_{(i)}$ is a commutative algebra, the $B(v)$ *projection algebra*, which we denote $A_{B(v)}$. The identity of the algebra is $I = \sum P_{(i)}$.

Definition 5. A subspace of $\mathbb{R}^k$ *respects* $B(v)$ if it has a basis in which each vector lies in some $V_{(i)}$.

For a subspace U we let $U^\perp$ be its orthogonal complement w.r.t. the standard inner product.

For U respecting $B(v)$ write $U = \operatorname{span}(\bigcup U_{(i)})$ for $U_{(i)} \subseteq V_{(i)}$. (Thus $U = \bigoplus U_{(i)}$ and $U_{(i)} = P_{(i)}U$.) Let $D_{(i)} = (U_{(i)})^\perp \cap V_{(i)}$. Then $(U_{(i)})^\perp = D_{(i)} \oplus \bigoplus_{j \ne i} V_{(j)}$.

Lemma 6. *A subspace $U^\perp$ respects $B(v)$ if U does.*

Proof. The subspaces of an inner product space form an orthocomplemented lattice in which the meet operation is intersection, and the negation operation is orthogonal complement. So for any subspaces W, W' we have De Morgan's law $(\operatorname{span}(W \cup W'))^\perp = W^\perp \cap W'^\perp$. Thus $U^\perp = \bigcap (U_{(i)})^\perp = \bigoplus D_{(i)}$. □

Lemma 7. *A subspace U respects $B(v)$ iff $U = \bigoplus (P_{(i)}U)$.*

Proof. ($\Leftarrow$): Because this gives an explicit representation of U as a direct sum of subspaces each restricted to some $V_{(i)}$.

($\Rightarrow$): By definition U is spanned by some collection of subspaces $V'_{(i)} \subseteq V_{(i)}$; since these subspaces are necessarily orthogonal, $U = \bigoplus V'_{(i)}$. Moreover, since $P_{(i)}$ annihilates $V_{(j)}, j \ne i$, and is the identity on $V_{(i)}$, it follows that each $V'_{(i)} = P_{(i)}U$. □

Theorem 8. *A subspace U is invariant under $v_{\bar{\odot}}$ iff U respects $B(v)$.*

Proof. ($\Leftarrow$): Let $w \in U$ and write $w = \sum w_i$ for $w_i \in U_{(i)}$. Then $v \odot w_i = \lambda_i w_i \in U_{(i)}$. So $v \odot w = \sum v \odot w_i \in \bigoplus U_{(i)} = U$.

($\Rightarrow$): If $U = v_{\bar{\odot}}(U)$ then these also equal $v_{\bar{\odot}}(v_{\bar{\odot}}(U))$, etc., so U is an invariant space of $A_{B(v)}$, meaning, $aU \subseteq U$ for any $a \in A_{B(v)}$. In particular, applying (1), this holds for $a = P_{(i)}$. So $U \supseteq \bigoplus (P_{(i)}U)$. On the other hand, since $\sum P_{(i)} = I$, $U = (\sum P_{(i)})U \subseteq \bigoplus (P_{(i)}U)$. So $U = \bigoplus (P_{(i)}U)$. Now apply Lemma 7. □

The symbol $\subset$ is reserved for strict inclusion.

Lemma 9. *If $R, T \subseteq [n]$ and $\operatorname{rowspace} \mathbb{H}(\mathbf{m}|_R) \subset \operatorname{rowspace} \mathbb{H}(\mathbf{m}|_{R \cup T})$, then there is a row $t \in T$ such that $\operatorname{rowspace} \mathbb{H}(\mathbf{m}|_R) \subset \operatorname{rowspace} \mathbb{H}(\mathbf{m}|_{R \cup \{t\}})$.*

Proof. Without loss of generality R, T are disjoint. Let $T' \subseteq T$ be a smallest set s.t. $\exists R' \subseteq R$ s.t. $\mathbf{m}_{R'} \odot \mathbf{m}_{T'} \notin \operatorname{rowspace} \mathbb{H}(\mathbf{m}|_R)$. Select any $t \in T'$ and write $\mathbf{m}_{R'} \odot \mathbf{m}_{T'} = \mathbf{m}_{R'} \odot \mathbf{m}_{T'-\{t\}} \odot \mathbf{m}_t$. By minimality of T', $\mathbf{m}_{R'} \odot \mathbf{m}_{T'-\{t\}} \in \operatorname{rowspace} \mathbb{H}(\mathbf{m}|_R)$. But then $\mathbf{m}_{R'} \odot \mathbf{m}_{T'} \in \operatorname{rowspace} \mathbb{H}(\mathbf{m}|_{R \cup \{t\}})$, so $\operatorname{rowspace} \mathbb{H}(\mathbf{m}|_R) \subset \operatorname{rowspace} \mathbb{H}(\mathbf{m}|_{R \cup \{t\}})$. □

Proof of Theorem 2. This is now a consequence of Lemma 9. Start with $R = \emptyset$, and repeatedly use the Lemma to adjoin to R a row from $[n] \setminus R$ which will increase the rank of $\mathbb{H}(\mathbf{m}|_R)$ by at least 1. □

Remark

$\operatorname{rank} \mathbb{H}(\mathbf{m})$, along with a basis (using only rows of $\mathbb{H}(\mathbf{m})$) for $\operatorname{rowspace} \mathbb{H}(\mathbf{m})$, can be computed in time $O(nk^3)$ using Chen and Moitra's "GrowByOne" procedure [3]. For completeness here is a version of that procedure: For $\ell \ge 0$ let $W_\ell = \operatorname{span}(\mathbf{m}|_{[\ell]})$, and let $r_\ell = \operatorname{rank} W_\ell$. W_ℓ is spanned by some vectors $\mathbf{m}_{S_{\ell,1}}, \ldots, \mathbf{m}_{S_{\ell,r_\ell}}$, with all $S_{\ell,i} \subseteq [\ell]$, which we compute as follows. For $\ell = 0$ we have $r_0 = 1$, $S_{0,1} = \emptyset$. For $\ell > 1$ form the matrix with rows $\mathbf{m}_{S_{\ell-1,1}}, \ldots, \mathbf{m}_{S_{\ell-1,r_{\ell-1}}}$ followed by rows $\mathbf{m}_{S_{\ell-1,1} \cup \{\ell\}}, \ldots, \mathbf{m}_{S_{\ell-1,r_{\ell-1}} \cup \{\ell\}}$. Perform Gaussian elimination to zero-out all but $r_\ell - r_{\ell-1}$ of the second batch of rows. The first batch, together with the non-eliminated rows of the second batch, become $\mathbf{m}_{S_{\ell,1}}, \ldots, \mathbf{m}_{S_{\ell,r_\ell}}$.

IV. Combinatorics of the NAE Condition: Proof of Theorem 4(a)

Recall we are to show: *If $\overline{\varepsilon}(\mathbf{m}) \ge -1$ then $\mathbf{m}$ has a restriction to some $k-1$ rows on which $\overline{\varepsilon} = -1$.*

Proof of Theorem 4(a). We induct on k. The (vacuous) base-case is $k = 1$.

For $k > 1$, we proceed by way of contradiction. Suppose the theorem fails for k, and let $\mathbf{m}$ be a k-column counterexample with the least possible number of rows, n. So $n > k - 1 \ge 1$. Necessarily every row of $\mathbf{m}$ is in $\operatorname{NAE}(\mathbf{m})$. Our strategy is to show $\mathbf{m}$ has a restriction $\mathbf{m}'$ to $n-1$ rows, for which $\overline{\varepsilon}(\mathbf{m}') \ge -1$; this will imply a contradiction because, by minimality of the number of rows of $\mathbf{m}$, $\mathbf{m}'$ has a restriction to $k-1$ rows on which $\overline{\varepsilon} = -1$.

If $\overline{\varepsilon}(\mathbf{m}) \ge 0$ then we can remove any single row of $\mathbf{m}$ and still satisfy $\overline{\varepsilon} \ge -1$.

Otherwise, $\overline{\varepsilon}(\mathbf{m}) = -1$, so there is a nonempty S such that $|\operatorname{NAE}(\mathbf{m}|^S)| = |S| - 1$; choose a largest such S. It cannot be that $S = [k]$ (as then $n = k - 1$). Arrange the rows $\operatorname{NAE}(\mathbf{m}|^S)$ as the bottom $|S| - 1$ rows of the matrix. As discussed earlier, for the NAE condition one may regard the distinct real values in each row of $\mathbf{m}$ simply as distinct colors; relabel the colors in each row above $\operatorname{NAE}(\mathbf{m}|^S)$ so the color above S is called "white." (There need be no consistency among the real numbers called white in different rows.) See Fig. 1.

Due to the maximality of $|S|$ and the fact that $\overline{\varepsilon}(\mathbf{m}) \ge -1$, there is no set of columns S' with $S \subset S'$ such that for some set of rows $A \subseteq [n] - \operatorname{NAE}(\mathbf{m}|^S)$, with $|A| = n - |S'| + 1$, $\mathbf{m}|_A^{S'}$ is all white. That is to say, if we form a bipartite graph

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3146630

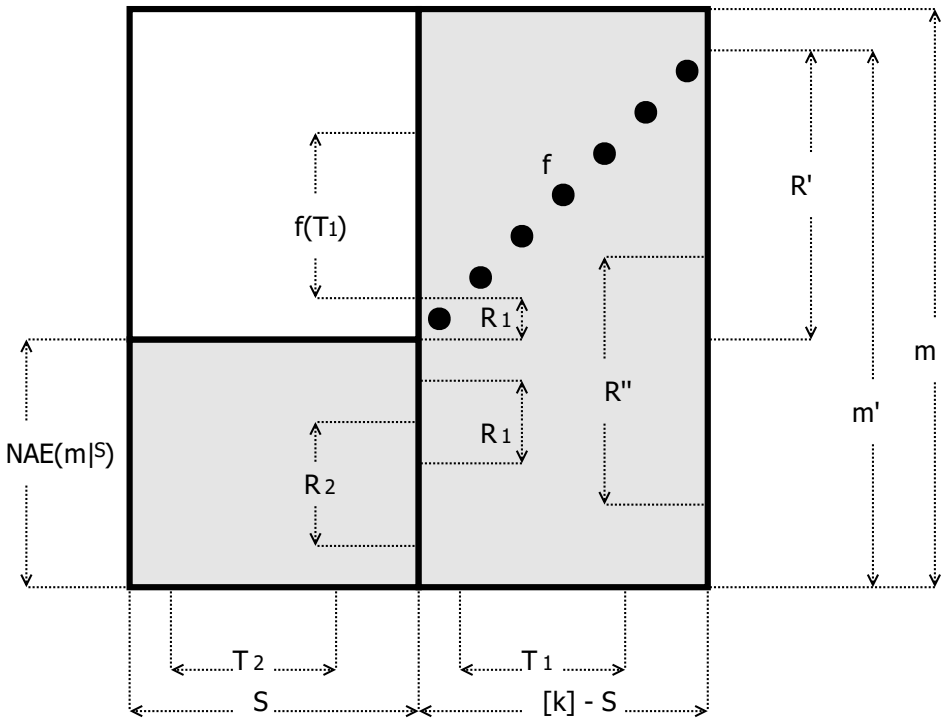


Fig. 1. Argument for Theorem 4(a). Upper-left region is white. Entries $(t, f(t))$ (indicated with black dots) are not white.

on "right" vertices corresponding to the columns $[k]-S$, and "left" vertices corresponding to the rows $[n]-\mathrm{NAE}(\mathbf{m}|^S)$, with non-white cells being edges, then any subset of the right vertices of size $\ell \geq 1$ has at least $\ell+1$ neighbors within the left vertices.

By the induction on k (since $S \neq \emptyset$), for the set of columns $[k]-S$ there is a set R'' of $k-|S|-1$ rows such that $\overline{\varepsilon}(\mathbf{m}|_{R''}^{[k]-S}) = -1$. Together with the rows of $\mathrm{NAE}(\mathbf{m}|^S)$ this amounts to at most $k-2$ rows, so since $n \geq k$, we can find two rows outside this union; delete either one of them, leaving a matrix $\mathbf{m}'$ with $n-1$ rows. This matrix has the rows $\mathrm{NAE}(\mathbf{m}|^S)$ at the bottom, and $n-|S|$ remaining rows which we call R'. The lemma will follow by showing that $\overline{\varepsilon}(\mathbf{m}') \geq -1$.

In $\mathbf{m}'$, the induced bipartite graph on right vertices $[k]-S$ and left vertices R' has the property that any right subset of size $\ell \geq 1$ has a neighborhood of size at least ℓ in R'. Applying Hall's Marriage Theorem, there is an injective $f : [k]-S \to R'$ employing only edges of the graph.

Now consider any nonempty set of columns T, and write it as $T = T_1 \cup T_2$ for $T_1 \subseteq [k]-S$ and $T_2 \subseteq S$. We need to show that $\varepsilon(\mathbf{m}'|^T) \geq -1$. Let $R_1 = \mathrm{NAE}(\mathbf{m}'|^{T_1}) \cap R''$ and $R_2 = \mathrm{NAE}(\mathbf{m}'|^{T_2})$. We have that $|R_1| \geq |T_1| - 1$ because $\overline{\varepsilon}(\mathbf{m}|_{R''}^{[k]-S}) = -1$. We further have that $|R_2| \geq |T_2| - 1$ because $\overline{\varepsilon}(\mathbf{m}) \geq -1$ and because $\mathrm{NAE}(\mathbf{m}|^{T_2}) \subseteq \mathrm{NAE}(\mathbf{m}|^S) = \mathrm{NAE}(\mathbf{m}'|^S)$, so no row of $\mathrm{NAE}(\mathbf{m}|^{T_2})$ has been removed in $\mathbf{m}'$.

If $T_2 = \emptyset$, the rows R_1 witness that $\varepsilon(\mathbf{m}'|^T) \geq -1$. Likewise if $T_1 = \emptyset$, the rows R_2 witness the same conclusion.

Lastly suppose both T_1 and T_2 are nonempty. Nonemptyness of T_2 gives $|\mathrm{NAE}(\mathbf{m}|^{T_2})| \geq |T_2| - 1$. Now use the matching f. The set of rows $f(T_1)$ lies in R' and is therefore disjoint from $\mathrm{NAE}(\mathbf{m}|^{T_2})$, which as noted is a subset of $\mathrm{NAE}(\mathbf{m}'|^S)$. Moreover since $T_2 \neq \emptyset$, every entry (t, j) for $t \in T_2, j \in R'$ is white. On the other hand due to the construction of f, for every $t \in T_1$ the entry $(t, f(t))$ is non-white. Therefore every row in $f(T_1)$ is in $\mathrm{NAE}(\mathbf{m}'|^{T_1 \cup T_2})$. So $|\mathrm{NAE}(\mathbf{m}'|^{T_1 \cup T_2})| \geq |T_2| - 1 + |T_1|$, which is to say $\varepsilon(\mathbf{m}'|^T) \geq -1$. Thus $\overline{\varepsilon}(\mathbf{m}') \geq -1$. □

V. From NAE to Rank: Proof of Theorem 4(b)

Recall we are to show: $\mathbb{H}(\mathbf{m})$ *has full column rank if* $\overline{\varepsilon}(\mathbf{m}) \geq -1$.

Proof of Theorem 4(b). The case $k = 1$ is trivial. Now suppose $k \geq 2$ and that Theorem 4(b) holds for all $k' < k$. Any constant rows of $\mathbf{m}$ affect neither the hypothesis nor the conclusion, so remove them, leaving $\mathbf{m}$ with at least $k-1$ rows. Now pick any set, C, of $k-1$ columns of $\mathbf{m}$. By Theorem 4(a) there are some $k-2$ rows of $\mathbf{m}$, call them R', on which $\overline{\varepsilon}(\mathbf{m}|_{R'}^C) = -1$. Let v be a row of $\mathbf{m}$ outside R'. Let R'' denote the set of rows of $\mathbf{m}$ other than v. Since R'' contains R', by induction $\dim \mathrm{rowspace}\, \mathbb{H}(\mathbf{m}|_{R''}^C) = k-1$. Therefore $U := \mathrm{rowspace}\, \mathbb{H}(\mathbf{m}|_{R''}) \subseteq \mathbb{R}^k$ is of dimension at least $k-1$. We claim now that $\dim v_{\bar{\odot}}(U) = k$. (Note that $v_{\bar{\odot}}(U) = \mathrm{rowspace}\, \mathbb{H}(\mathbf{m})$.)

Suppose to the contrary that $\dim v_{\bar{\odot}}(U) = k-1$. It must then be that $\dim U = k-1$ and $v_{\bar{\odot}}(U) = U$. So as proven in Theorem 8, U respects $B(v)$. Since v is nonconstant, $B(v)$ is a partition of $[k]$ into $\ell \geq 2$ nonempty blocks $B(v)_{(i)}$, and $U = \bigoplus_{i=1}^{\ell} U_{(i)}$ with $U_{(i)} = P_{(i)} U_{(i)}$. So there is some i_0 for which $U_{(i_0)} \subset V_{(i_0)}$; specifically, $U_{(i)} = V_{(i)}$ for all $i \neq i_0$, and $\dim U_{(i_0)} = \dim V_{(i_0)} - 1$. Since $|B(v)_{(i_0)}| < k$, we know by induction that $P_{(i_0)}\, \mathrm{rowspace}\, \mathbb{H}(\mathbf{m}) = V_{(i_0)}$.

But since $\operatorname{rowspace}\mathbb{H}(\mathbf{m}) = v_{\bar{\odot}}(U) = U$, this means that $P_{(i_0)}U = V_{(i_0)}$. Contradiction. □

ACKNOWLEDGMENT

We thank the anonymous referees for their careful review which substantially improved the manuscript.

REFERENCES

[1] A. Anandkumar, D. J. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In *Proc. 25th Ann. Conf. on Learning Theory - COLT*, volume 23 of *JMLR Proceedings*, pages 33.1–33.34, 2012. URL: http://proceedings.mlr.press/v23/anandkumar12/anandkumar12.pdf.

[2] K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In *Proc. 21st Ann. Conf. on Learning Theory - COLT*, pages 9–20. Omnipress, 2008. URL: http://colt2008.cs.helsinki.fi/papers/7-Chaudhuri.pdf.

[3] S. Chen and A. Moitra. Beyond the low-degree algorithm: mixtures of subcubes and their applications. In *Proc. 51st Ann. ACM Symp. on Theory of Computing*, pages 869–880, 2019. doi:10.1145/3313276.3316375.

[4] M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM J. Comput.*, 31(2):375–397, 2001. doi:10.1137/S0097539798342496.

[5] B. S. Everitt and D. J. Hand. *Mixtures of discrete distributions*, pages 89–105. Springer Netherlands, Dordrecht, 1981.

[6] J. Feldman, R. O'Donnell, and R. A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM J. Comput.*, 37(5):1536–1564, 2008. doi:10.1137/060670705.

[7] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proc. 12th Ann. Conf. on Computational Learning Theory*, pages 53–62, July 1999. doi:10.1145/307400.307412.

[8] S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman. Source identification for mixtures of product distributions. In *Proc. 34th Ann. Conf. on Learning Theory - COLT*, volume 134 of *Proceedings of Machine Learning Research*, pages 2193–2216. PMLR, 2021. URL: http://proceedings.mlr.press/v134/gordon21a.html.

[9] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, and L. Sellie. On the learnability of discrete distributions. In *Proc. 26th Ann. ACM Symp. on Theory of Computing*, pages 273–282, 1994. doi:10.1145/195058.195155.

[10] J. Li, Y. Rabani, L. J. Schulman, and C. Swamy. Learning arbitrary statistical mixtures of discrete distributions. In *Proc. 47th Ann. ACM Symp. on Theory of Computing*, pages 743–752, 2015. doi:10.1145/2746539.2746584.

[11] B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*, volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, 1995.

[12] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, 6(1):355–378, 2019. doi:10.1146/annurev-statistics-031017-100325.

[13] S. Newcomb. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4):343–366, 1886.

[14] K. Pearson. Contributions to the mathematical theory of evolution III. *Philosophical Transactions of the Royal Society of London (A.)*, 185:71–110, 1894.

[15] Y. Rabani, L. J. Schulman, and C. Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *Proc. 5th Conf. on Innovations in Theoretical Computer Science*, pages 207–224, 2014. doi:10.1145/2554797.2554818.

[16] B. Tahmasebi, S. A. Motahari, and M. A. Maddah-Ali. On the identifiability of finite mixtures of finite product measures. (Also in "On the identifiability of parameters in the population stratification problem: A worst-case analysis," Proc. ISIT 2018 pp. 1051-1055), 2018. URL: https://arxiv.org/abs/1807.05444.

[17] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Inc., 1985.

Spencer L. Gordon received the B.Sc. in Computer Science from Brown University in 2014 and the M.Sc. in Computer Science from the University of Illinois at Urbana-Champaign in 2017. He is currently a Ph.D. candidate at the California Institute of Technology.

Leonard J. Schulman received the B.Sc. in Mathematics in 1988 and the Ph.D. in Applied Mathematics in 1992, both from the Massachusetts Institute of Technology. Since 2000 he has been on the faculty of the California Institute of Technology. He has also held appointments at UC Berkeley, the Weizmann Institute of Science, the Georgia Institute of Technology, and the Israel Institute for Advanced Studies.