Hadamard Extensions and the Identification of Mixtures of Product Distributions

Spencer L. Gordon∗ Leonard J. Schulman†

February 16, 2021

Abstract

The Hadamard Extension of a matrix is the matrix consisting of all Hadamard products of subsets of its rows. This construction arises in the context of identifying a mixture of product distributions on binary random variables: full column rank of such extensions is a necessary ingredient of identification algorithms. We provide several results concerning when a Hadamard Extension has full column rank.

1 Introduction

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

\[
\begin{align*}
u \circ v := (u_1 v_1, \ldots, u_k v_k)
\end{align*}
\]

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

\[
\begin{align*}
u \cdot v \circ = v \circ u.
\end{align*}
\]

Throughout this paper \( m \) is a real matrix with row set \( [n] := \{1, \ldots, n\} \) and column set \( [k] \); write \( m_i \) for a row and \( m^j \) 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 columns and rows (with either index omitted if all rows or columns are retained).

**Definition 1.** The Hadamard Extension of \( m \), written \( \mathcal{H}(m) \), is the \( 2^n \times k \) matrix with rows \( m_S \) for all \( S \subseteq [n] \); where, for \( S = \{i_1, \ldots, i_\ell\} \), \( m_S = m_{i_1} \circ \cdots \circ m_{i_\ell} \); equivalently \( m_S^j = \prod_{i \in S} m_i^j \). (In particular \( m_\emptyset = 1 \).)

This construction has arisen recently in learning theory \cite{3,8} where it is essential to source identification for a mixture of product distributions on binary random variables. We explain the connection further in Section 5.

Motivated by this application, we are interested in the following two questions:

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

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

In answer to the first question we show in Section 2:

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

∗Engineering and Applied Science, California Institute of Technology, slgordon@caltech.edu.
†Engineering and Applied Science, California Institute of Technology, schulman@caltech.edu. Research supported in part by NSF grant CCF-1909972.
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 it cannot be full rank. Extending this further, suppose there are three columns \( C \) in which only one row \( r \) has more than one color. Then Rowspace \( \mathbb{H}(\mathbf{m}[^C]) \) is spanned by \( 1[^C] \) and \( r[^C] \), so again \( \mathbb{H}(\mathbf{m}) \) cannot be full rank. Motivated by these necessary conditions, set:

**Definition 3.** For a matrix \( Q \) let NAE\((Q)\) be the set of nonconstant rows of \( Q \) (NAE = “not all equal”); let \( \varepsilon(Q[^C]) = |\text{NAE}(Q[^C])| - |C| \) and let \( \tau(Q) = \min_{C \neq \emptyset} \varepsilon(Q[^C]) \). If \( \tau(Q) \geq -1 \) 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

1. There is a restriction of \( \mathbf{m} \) to some \( k - 1 \) rows \( R \) such that \( \tau(\mathbf{m}|_R) = -1 \).
2. \( \mathbb{H}(\mathbf{m}) \) 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 \( \mathbf{m}_1, \mathbf{m}_1[^1], \mathbf{m}_1[^2], \ldots, \mathbf{m}_1[^{k-1}] \) 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 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 \( \ell \) columns \( C \) s.t. \( \varepsilon(\mathbf{m}[^C]) = -1 \).

For \( k > 3 \) the NAE condition is no longer necessary for \( \mathbb{H}(\mathbf{m}) \) to have full column rank. E.g., for \( k = 2^\ell \), the \( \ell \times k \) “Hamming matrix” \( \mathbf{m} = (-1)^{j_i} \) where \( j \) is a \( \ell \)-bit string \( j = (j_1, \ldots, j_\ell) \), forms \( \mathbb{H}(\mathbf{m}) \) = the Fourier transform for the group \((\mathbb{Z}/2)^\ell \), (often called a Hadamard matrix), which is invertible. Furthermore, almost all (in the sense of Lebesgue measure) \(|\log k| \times k \) matrices \( \mathbf{m} \) form a full-rank \( \mathbb{H}(\mathbf{m}) \). (This is because \( \det \mathbb{H}(\mathbf{m}) \) is a polynomial in the entries of \( \mathbf{m} \), and the previous example shows the polynomial is nonzero.) 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.

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

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

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

and introduce the notation

\[
v_\odot(U) = \text{span}\{U \cup v_\odot(U)\}.
\]

We want to understand which subspaces \( U \) are invariant under \( v_\odot \). Let \( v \) have distinct values \( \lambda_1 > \ldots > \lambda_\ell \) for \( \ell \leq k \). 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. standard inner product. We have the matrix equation

\[
p_{v,i}(v_\odot) = P(i).
\]

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 is spanned by vectors each of which lies in some \( V(i) \).

For \( U \) respecting \( B(v) \) write \( U = \text{span}(U) \) for \( U(i) \subseteq V(i). \) Let \( D(i) = (U(i))^\perp \cap V(i) \). Then \( (U(i))^\perp = D(i) \oplus \bigoplus_{j \neq i} V(j) \).
Lemma 6. Subspace $U^\perp$ respects $B(v)$ if $U$ does.

Proof. In general, $(\text{span}(W \cup W'))^\perp = W^\perp \cap W'^\perp$. So $U^\perp = \bigcap_i (U_i)^\perp = \bigoplus_i D_i$. □

Lemma 7. 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 $V_{(j)}$, $i \neq j$, and is the identity on $V_i$, it follows that each $V_i' = P_i U$.

Theorem 8. Subspace $U$ is invariant under $v_\otimes$ iff $U$ respects $B(v)$.

Proof. $(\Leftarrow)$: It suffices to show $U^\perp$ is invariant under $v_\otimes$. By the previous lemma, it is equivalent to suppose that $U^\perp$ respects $B(v)$. So let $d \in U^\perp$ and write $d = \sum_i d_i \in D_i$. Then $v \otimes d_i = \lambda_i d_i \in D_i$. So $v \otimes d = \sum v \otimes d_i \in \bigoplus D_i = U^\perp$.

$(\Rightarrow)$: If $U = v_\otimes(U)$ then these also equal $v_\otimes(v_\otimes(U))$, etc., so $U$ is an invariant space of $A_{B(v)}$, meaning, $aU \subseteq U$ for any $a \in A_{B(v)}$. In particular 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 $S, T \subseteq [n]$ and Rowspace $\mathbb{H}(m|S) \subset$ Rowspace $\mathbb{H}(m|S \cup T)$, then there is a row $t \in T$ such that Rowspace $\mathbb{H}(m|S) \subset$ Rowspace $\mathbb{H}(m|S \cup \{t\})$.

Proof. Without loss of generality $S, T$ are disjoint. Let $T' \subseteq T$ be a smallest set s.t. $\exists S' \subseteq S$ s.t. $m_{S'} \circ m_{T'} \notin$ Rowspace $\mathbb{H}(m|S)$. Select any $t \in T'$ and write $m_{S'} \circ m_{T'} = m_{S'} \circ m_{T'-\{t\}} \circ m_t$. By minimality of $T'$, $m_{S'} \circ m_{T'-\{t\}} \in$ Rowspace $\mathbb{H}(m|S)$. But then $m_{S'} \circ m_{T'} \in$ Rowspace $\mathbb{H}(m|S \cup \{t\})$, so Rowspace $\mathbb{H}(m|S) \subset$ Rowspace $\mathbb{H}(m|S \cup \{t\})$.

Theorem 2 is now a consequence of Lemma 9.

It follows from Theorem 2 that we can check whether rank $\mathbb{H}(m) = k$ in time $O(n)^k$ by computing rank $\mathbb{H}(m|S)$ for each $S \in \binom{[n]}{k-1}$.

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

Recall we are to show: 4(a). If $\tau(m) \geq -1$ then $m$ has a restriction to some $k-1$ rows on which $\tau = -1$.

Proof. We induct on $k$. The (vacuous) base-case is $k = 1$. For $k > 1$, we induct on $n$, with base-case $n = k - 1$.

Supposing the Theorem fails for $k, k > 1$, let $m$ be a $k$-column counterexample with least $n$. Necessarily every row is in NAE($m$), and $n > k - 1 \geq 1$. We will show $m$ has a restriction $m'$ to $n - 1$ rows, for which $\tau(m') \geq -1$; this will imply a contradiction because, by minimality of $m$, $m'$ has a restriction to $k - 1$ rows on which $\tau = -1$.

If $\tau(m) \geq 0$ then we can remove any single row of $m$ and still satisfy $\tau \geq -1$.

Otherwise, $\tau(m) = -1$, so there is a nonempty $S$ such that $|\text{NAE}(m^S)| = |S| - 1$; choose a largest such $S$. It cannot be that $S = [k]$ (as then $n = k - 1$). Arrange the rows NAE($m^S$) 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 $m$ simply as distinct colors; relabel the colors in each row above NAE($m^S$) so the color above $S$ is called “white.” (There need be no consistency among the real numbers called white in different rows.) See Figure 4.

Due to the maximality of $|S|$, there is no white rectangle on $\ell$ columns and $n - |S| - \ell + 1$ rows inside $m^{|S| - \ell} \cap \text{NAE}(m^S)$ for any $\ell \geq 1$. That is to say, if we form a bipartite graph on right vertices corresponding to the columns $[k] - S$, and left vertices corresponding to the rows $[n] - \text{NAE}(m^S)$, 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.
Recall we are to show: \(4(\text{b})\) \(H(m)\) has full column rank if \(\pi(m) \geq -1\).

**Proof.** The case \(k = 1\) is trivial. Now suppose \(k \geq 2\) and that Theorem \(4(\text{b})\) holds for all \(k' < k\). Any constant rows of \(m\) affect neither the hypothesis nor the conclusion, so remove them, leaving \(m\) with at least \(k - 1\) rows. Now pick any set, \(C\), of \(k - 1\) columns of \(m\). By Theorem \(4(\text{a})\) there are some \(k - 2\) rows of \(m\), call them \(R'\), on which \(\pi(m|_{R'}) = -1\). Let \(v\) be a row of \(m\) outside \(R'\). Call the rows of \(m\) apart from \(v\), \(R''\). Since \(R''\) contains

---

**Figure 1:** Argument for Theorem \(4(\text{a})\). Upper-left region is white. Entries \((t, f(t))\) are not white.
Consider observable random variables $X_1, \ldots, X_n$ that are statistically independent conditional on $H$, a hidden random variable supported on $\{1, \ldots, k\}$. (See causal diagram.)

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

Finite mixture models were 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 finite mixture models were pioneered in the late 1800s in [13, 14]. The problem of learning such distributions parameters are $w$ in fact $U$. Since $\dim U = k$. (Further detail for the last step: let $w\not\in U$ s.t. $\dim U = V(i_0)$. Thus in fact $U = \mathbb{R}^k$. (Further detail for the last step: let $w\not\in U$ s.t. $\dim U = V(i_0)$. Then $w = (I - P(i_0))(w - w')$. Then $w + w'' \in \text{Rowspace } \mathbb{H}(m)$, and $w' + w'' = w$.)

5 Motivation

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

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

Finite mixture models were 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 finite mixture models were pioneered in the late 1800s in [13, 14]. The problem of learning such distributions parameters are $w$ in fact $U$. Since $\dim U = k$. (Further detail for the last step: let $w\not\in U$ s.t. $\dim U = V(i_0)$. Then $w = (I - P(i_0))(w - w')$. Then $w' + w'' \in \text{Rowspace } \mathbb{H}(m)$, and $w' + w'' = w$.)

Connection to rank $\mathbb{H}(m)$. In general $\mu$ is not injective (even allowing for permutation among the values of $H$). For instance it is clearly not injective if $m$ has two identical columns (unless $\pi$ places no weight on those). More generally, and assuming all $\pi_j > 0$, it cannot be injective unless $\mathbb{H}(m)$ has full column rank.

One sufficient condition for injectivity, due to [16], is that there be $2k - 1$ “separated” observables $X_i$; $X_i$ is separated if all $m^j_i$ are distinct, or in our terminology, if no color recurs in $m_i$. (Further [8], one can lower bound the distance between $\mu(m, \pi)$ and any $\mu(m', \pi')$ in terms of $\min_{i,j} |m^j_i - m'^j_i|$ and the distance between $(m, \pi)$ and $(m', \pi)$.)

A weaker sufficient condition for injectivity of $\mu$, due to [8], is that for every $i \in [n]$ there exist two disjoint sets $A, B \subseteq [n] \setminus \{i\}$ such that $\mathbb{H}(m|A)$ and $\mathbb{H}(m|B)$ have full column rank. (It is not known whether two disjoint such $A, B$ are strictly necessary, but the implied $n \leq 2k - 1$ is in general best possible [15].)
References

[1] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden Markov models. In Proc. 25th Ann. Conf. on Computational Learning Theory, pages 33.1–33.34, 2012.

[2] K. Chaudhuri and S. Rao. Learning mixtures of product distributions using correlations and independence. In Proc. 21st Ann. Conf. on Computational Learning Theory, pages 9–20, 2008.

[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.

[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. Prev. FOCS ’98.

[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.

[7] Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In Proc. 12th Ann. Conf. on Computational Learning Theory, pages 183–192, July 1999.

[8] S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman. Source identification for mixtures of product distributions. arXiv:2012.14540, 2020.

[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.

[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.

[11] B. G. Lindsay. Mixture models: theory, geometry and applications. In NSF-CBMS regional conference series in probability and statistics, pages i–163. JSTOR, 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.

[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.

[16] B. Tahmasebi, S. A. Motahari, and M. A. Maddah-Ali. On the identifiability of finite mixtures of finite product measures. IEEE International Symposium on Information Theory (ISIT) 2018 and arXiv:1807.05444v1, 2018.

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