Source Identification for Mixtures of Product Distributions

Spencer L. Gordon SLGORDON@CALTECH.EDU
Engineering and Applied Science, California Institute of Technology, Pasadena CA 91125, USA

Bijan Mazaheri BMAZAHER@CALTECH.EDU
Engineering and Applied Science, California Institute of Technology, Pasadena CA 91125, USA

Yuval Rabani YRABANI@CS.HUJI.AC.IL
The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 9190416, Israel

Leonard J. Schulman SCHULMAN@CALTECH.EDU
Engineering and Applied Science, California Institute of Technology, Pasadena CA 91125, USA

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We give an algorithm for source identification of a mixture of k product distributions on n bits. This is a fundamental problem in machine learning with many applications. Our algorithm identifies the source parameters of an identifiable mixture, given, as input, approximate values of multilinear moments (derived, for instance, from a sufficiently large sample), using $2^{O(k^2)}n^{O(k)}$ arithmetic operations. Our result is the first explicit bound on the computational complexity of source identification of such mixtures. The running time improves previous results by Feldman, O'Donnell, and Servedio (FOCS 2005) and Chen and Moitra (STOC 2019) that guaranteed only learning the mixture (without parametric identification of the source). Our analysis gives a quantitative version of a qualitative characterization of identifiable sources that is due to Tahmasebi, Motahari, and Maddah-Ali (ISIT 2018).

1. Introduction

1.1. The problem

Consider observable random variables $X_1, \ldots, X_n$ that are distributed on a common range R. In a finite mixture model, the joint distribution on these random variables is governed by a *hidden* or *latent* random variable H supported on $\{1, \ldots, k\}$, such that $X_1, \ldots, X_n$ are statistically independent conditional on H. We consider the case of a finite range R. The hardest and most fundamental case is when the range is binary (i.e., $R = \{0, 1\}$) and there are no further constraints relating the distributions at the observables; the case of larger R reduces to this case (see Feldman et al. (2008), for example). We will refer to observable random variables with binary range as "observable bits."

This paper provides a novel algorithm that identifies the source parameters of a mixture of k product distributions on $n \geq 3k - 3$ observable bits. The algorithm requires that at least $3k - 3$ of the observable bits are ζ-separated (definition below), a sufficient condition for identifiability.

The inputs to our algorithm are empirical estimates of the "multilinear moments" $E(X_S)$ where $X_S = \prod_{i \in S} X_i$ and $|S| \leq 3k - 3$. The expectation of a Bernoulli random variable can be estimated to within an additive factor of ε with sample size roughly $1/\varepsilon^2$. Hence, we will assume our empirical moments deviate from their true values by an error of at most $\varepsilon < \zeta^{O(k^2 \log k)}$.

The algorithm identifies the parameters of the mixture model to an accuracy of $\zeta^{-O(k^2 \log k)}\varepsilon$. This is also roughly (up to a different constant hidden by the big-Oh notation) the statistical distance between the empirical distribution and the output (i.e., the learned) distribution. The runtime of our algorithm is $2^{O(k^2)}n^{O(k)}$ arithmetic operations.[1] [2] Under a stronger assumption that *all* observable bits are ζ-separated, the runtime improves to $2^{O(k^2)}n$.

Our runtime of $2^{O(k^2)}n^{O(k)}$ improves on the best previous $k^{O(k^3)}n^{O(k^2)}$;[3] for n extremely large relative to k this is an improvement from $n^{O(k^2)}$ to $n^{O(k)}$. The most interesting algorithmic and statistical aspects of the problem, however, are best brought out by considering k as the primary parameter; from this perspective (by taking n subexponential in k), it is then evident that our principal contribution is a runtime improvement from $\exp((1+o(1))k^3)$ to $\exp((1+o(1))k^2)$.

1.2. Mixture models

Finite mixture models were pioneered in the late 1800s in Newcomb (1886); Pearson (1894) in the context of applications in astronomy and the mathematical theory of evolution. It is difficult to do justice to the vast literature in statistics on mixture models; see, e.g., the surveys Everitt and Hand (1981); Titterington et al. (1985); Lindsay (1995); McLachlan et al. (2019). The computational complexity of learning mixture models was studied starting with the seminal papers Kearns et al. (1994); Cryan et al. (2001); Dasgupta (1999); Freund and Mansour (1999). The machine learning community has shown recent interest in learning mixtures of product distributions for pattern recognition. Motivating applications abound in population genetics, bioinformatics, image recognition, text classification, and other areas, e.g., Pritchard et al. (2000); Ji et al. (2005); Juan and Vidal (2004); Juan and Vidal (2002). The algorithms in this context are primarily based on the method of iterative Expectation Maximization (EM) clustering (e.g., Juan et al. (2004); Li et al. (2016); Palmer et al. (2016); Carreira-Perpiñán and Renals (2000)). As detailed further below, in this genre provable guarantees of source identification are provided in Najafi et al. (2020), but the runtime depends exponentially on the sample size, and a very large n is required. In the theory of computing literature, special cases and variants of learning mixtures of product distributions were considered in Cryan et al. (2001); Freund and Mansour (1999); Chaudhuri and Rao (2008); Arora et al. (2012); Anandkumar et al. (2012a,b); Rabani et al. (2014); Li et al. (2015); Kim et al. (2019); Chen and Moitra (2019); Gordon et al. (2020). A more detailed account on the previous results Feldman et al. (2008); Chen and Moitra (2019) on mixture learning is given below.

1.3. Identification (parameter estimation) vs. learning

The question of *source identification* dates back in the statistics literature at least to 1950 Koopmans and Reiersol (1950); Koopmans (1950); Teicher (1963); Blischke (1964); Yakowitz and Spragins (1968). This is still a thriving area of research (see, e.g., Carreira-Perpiñán and Renals (2000); Allman et al. (2009); Vandermeulen and Scott (2015); Tahmasebi et al. (2018); Ritchie et al. (2020); Aragam et al. (2020)). From a computational perspective, the "learning" vs. "source identification"

1. We note that the runtime relates to the post-sampling computation, after aggregating the empirically observed frequencies. There are good reasons for making this distinction. Collecting the sample and computing the frequencies is computationally trivial. It can often be done under a streaming model or in parallel. Or the frequencies might be available from an external source.
2. We suppress here dependence on the mixture weights.
3. This previous work actually focused on the "learning" problem; see Section 1.3. We suppress here dependence on the learning accuracy.

contrast was raised in Freund and Mansour (1999), in the context that interests us here of learning a mixture of several (in that work, two) binary product distributions. "Learning" means computing any hypothetical model that generates observable statistics close to the empirical ones. "Source identification" means computing a model that is close in parameter space to the true model underlying the empirical statistics. Clearly, source identification is a more challenging goal that implies also learning. Source identification is desirable for a variety of reasons. A key reason is the danger of overfitting a model to the empirical data, thereby ruining its predictive guarantees (see, for instance, Koller and Friedman (2009) ch. 16). Another reason is that source identification is necessary for the quantification of causal relations between the hidden variable and the observable variables (see Pearl (2009); Spirtes et al. (2000) for an introduction to graphical causal models)—here identification, but not learning, provides explanatory value along with the possibility of effective intervention. Finally, whether in the causal context or otherwise, identification provides the statistician with *semantics*—an actual characterization of the process generating the data—which may be far more useful than the mere ability to artificially generate samples from the same distribution.

The distinction between learning and source identification has two main aspects. First, a distribution on observable variables might be inherently explainable by two (or more) far-apart models, ruling out source identification. Thus, we must restrict our attention to a class of *identifiable* models, for which the mapping of model to distribution on observables is one-to-one. In this paper, we consider the case that there are at least $2k-1$ observables that are ζ-separated. An observable bit X is *separated* if the k conditional probabilities $\Pr[X=1 \mid H=j]$, $j=1,\ldots,k$ are mutually distinct, and ζ-*separated* if every two values differ by at least $\zeta > 0$. (Necessarily $\zeta \leq \frac{1}{k-1}$.) Having $2k-1$ observables that are separated is a sufficient condition for identification Tahmasebi et al. (2018). [4]

Second, even if a model is identifiable in the limit of perfect statistics (infinite sample size), the available empirical statistics might be insufficiently accurate, yet still allow for the learning objective. It is obvious, though, that a learning algorithm that runs on data produced by an *identifiable* source and is required to achieve sufficiently high accuracy also implicitly identifies the source.[5] However, a learning algorithm might learn a mixture model without this implying that the empirical data enables identification. In fact, the ground-breaking work of Feldman et al. (2008) gives an $(nk/\varepsilon)^{O(k^3)}$ time algorithm for *learning* a model of k-mixture of n binary product distributions that generates a distribution on the observables within statistical distance ε of the empirical statistics. A faster $k^{O(k^3)}(n/\varepsilon)^{O(k^2)}$ time learning algorithm is given in Chen and Moitra (2019).[6] The algorithms in these papers are not guaranteed, under any assumptions, to identify the source to any particular accuracy ε; in fact the algorithms succeed even in cases where the source is not identifiable.

1.4. Our work

Our contribution is twofold. First, we prove a quantitative version of the identifiability criterion: namely, for any given $\zeta > 0$, we establish (roughly; see later for exact statements) that any two models which differ by η in parameter space, differ in their statistics by at least $\eta\zeta^{O(k^2 \log k)}$. [7] This,

4. Depending on the parameters, fewer sources may suffice. But in some cases $2k-1$ are necessary, including the important case when the X_i are iid conditional on H, which we call the "power distribution" case Rabani et al. (2014), since terms of the form $\prod_{i=1}^{n} \Pr[X_i = 1 \mid H = j]$ are replaced by $(\Pr[X_1 = 1 \mid H = j])^n$.
5. This is because the space of source parameters is compact and the mapping to observable statistics is continuous.
6. Notice, in particular, footnote 2 in that paper.
7. To simplify the informal discussion, we state most of the bounds in the rest of this paragraph just for $\eta \geq \zeta^{O(k^2 \log k)}$.

of course, implies that the algorithms of Feldman et al. (2008); Chen and Moitra (2019) can be used for source identification, assuming ζ-separation and sufficiently small target accuracy $\varepsilon = \zeta^{O(k^2 \log k)}$. Second, we improve substantially over the runtime of these two algorithms, so that (even under the conditions under which our result imples that those algorithms can perform identification), our source identification algorithm is more efficient. Specifically, if we have at least $3k-3$ observables that are ζ-separated, our algorithm requires empirical statistics accuracy $\zeta^{O(k^2 \log k)}$ (i.e., sample size $\zeta^{-O(k^2 \log k)}$) and has a runtime of $2^{O(k^2)} n^{O(k)}$ arithmetic operations. Our algorithm can also identify the source using the minimum of $2k-1$ ζ-separated observables; but then we require input accuracy $\zeta^{O(k^3)}$ (but the same $2^{O(k^2)} n^{O(k)}$ runtime). If *all* observables are ζ-separated, then the runtime improves to[8] $2^{O(k^2)} n$ (with the same input accuracies, according to the number of variables). Our contributions establish quantitative bounds on the qualitative sufficiency of ζ-separation of Tahmasebi et al. (2018); the results in that paper are entirely non-algorithmic. The only explicit algorithmic result on source identification that we are aware of is Najafi et al. (2020). Their algorithm, which runs under a somewhat more general assumption than ζ-separation, requires complete enumeration over the choice of mixture constituent that generated each sample point, for a sufficiently large sample. Thus, it is prohibitively expensive, requiring at least $n = \exp(k^2)$ observable random variables (as compared with $3k-3$), and runtime that is doubly exponential, namely $k^{\exp(k^2)}$.

1.5. The new method: bootstrapping synthetic bits

The main idea underlying our algorithm is the following. Given sufficiently many ζ-separated observables, for which we have sufficiently accurate empirical multilinear moments,[9] we show how to construct "synthetic bits" for which we can compute highly accurate power moments, i.e., the moments that occur in the far more restricted problem of power distributions. The higher moments of these synthetic bits are created out of linear combinations of multilinear moments of the original bits. This mechanism in its idealized form (i.e., for perfect statistics) suffices to re-prove the theorem of Tahmasebi et al. (2018) that $2k-1$ separated observables suffice for source identification. The next challenge we face is to bound the coefficients of the multilinear monomials in these linear combinations, as this affects the required accuracy of the empirical statistics (thus, the required sample size). The synthetic bits method reduces the problem to the special case of identification of a mixture of k power distributions, i.e., when the X_i-s are iid conditional on H. This special case is effectively an extension of the theory of orthogonal polynomials on the reals and the classical *moment problem* Schmüdgen (2017); Simon (2015), and methods such as Prony's method or the Matrix Pencil method were shown to solve it Rabani et al. (2014); Li et al. (2015); Kim et al. (2019); Gordon et al. (2020). Despite the power distributions case being so highly constrained, it has useful applications, e.g., in reconstructing population histories and in learning topic models (see the above references). We can use these existing algorithms to recover the mixture of power distributions on synthetic bits. That in turn enables recovery of the mixture of product distributions on the observable bits. The best algorithm for power distributions to date Gordon et al. (2020) requires estimates of the first $2k$ moments of the synthetic bits to within accuracy $\zeta^{O(k)}$, and has runtime $O(k^{2+o(1)})$. Thus, this component of our runtime is cheap; the runtime bound of our algorithm is dominated by the

8. Here we suppress a $O(\log\log(\varepsilon^{-1}))$ term.
9. We call these empirical moments as we expect them to obtained by sampling; but our theorems depend only on their being sufficiently accurate.

exhaustive search for $3k-3$ observables that are ζ-separated (unless all observables are known to be ζ-separated), and by the construction of the $2k$ synthetic bits.

2. Preliminaries and main theorem

Notation There is a hidden variable H ranging in $[k]$, and n binary observable variables X_i; write $\mathbf{m}_{ij} := \Pr[X_i = 1 \mid H = j]$. The distribution of the hidden variable is denoted $\pi_j := \Pr(H = j)$. The model parameters are thus $(\pi_j)_{j\in[k]}$ and $(\mathbf{m}_{ij})_{i\in[n],j\in[k]}$ (up to permuting $[k]$).

Vectors are row vectors unless otherwise indicated. For $S \subseteq [n]$ define the random variable $X_S = \prod_{i\in S} X_i$. We make extensive use of *Hadamard product* for vectors $u = (u_1, \ldots, u_k)$, $v = (v_1, \ldots, v_k)$:

$$\odot : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k \tag{1}$$

$$u \odot v = (u_1 v_1, \ldots, u_k v_k) \tag{2}$$

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

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

Let $\mathbf{m}_i$ be the row vector $(\mathbf{m}_{i1}, \ldots, \mathbf{m}_{ik})$. Following Chen and Moitra (2019) here and in (4), let $\mathbf{M} \in \mathbb{R}^{2^{[n]}\times k}$ be the matrix with rows indexed by subsets $S = \{i_1, \ldots, i_s\} \subseteq [n]$, with rows

$$\mathbf{M}_S = \mathbf{m}_{i_1} \odot \mathbf{m}_{i_2} \odot \cdots \odot \mathbf{m}_{i_s}. \tag{3}$$

In particular, $\mathbf{M}_\emptyset = \mathbb{1}$ and $\mathbf{M}_{\{i\}} = \mathbf{m}_i$ for all $i \in [m]$.

Observe that source identification is not possible if $\mathbf{M}$ has less than full column rank, i.e., $\mathrm{rank}\,\mathbf{M} < k$, as then the mixing weights cannot be unique.

For a collection of subsets $\mathcal{S} \subseteq 2^{[n]}$, let $\mathbf{M}[\mathcal{S}]$ denote the restriction of $\mathbf{M}$ to the rows $\mathbf{M}_S$, $S \in \mathcal{S}$. E.g., $\mathbf{M} = \mathbf{M}[2^{[n]}]$.

The empirical multi-linear moments For a finite sample drawn from the model, we let $\tilde{\mathrm{g}}(S)$ be the empirical estimate of $\mathbb{E}[X_S]$, i.e., the fraction of samples for which $\prod_{i\in S} X_i = 1$. These $\tilde{\mathrm{g}}(S)$ for $S \subseteq [n]$ are the complete list of "observables" of the model. Each converges, in the infinite-sample limit, to the value $\mathrm{g}(S) := \mathbb{E}[X_S]$,

$$\mathrm{g}(S) = \mathbf{M}_S \pi^{\mathsf{T}} = \mathbf{M}_S \pi_\odot \mathbb{1}^{\mathsf{T}}.$$

Comparison to the iid case The fact that we have access *only* to multi-linear moments is the key constraint of the problem. Compare with the "power" case, i.e., when we know in advance that all rows $\mathbf{m}_i$ are identical. (That is, the X_i are iid conditional on H.) Then the statistics of X_2 tell us nothing about rv X_1. But the statistics of X_1X_2 do tell us something new about X_1: X_1X_2 is distributed as would be a binary observable whose row had entries $(\mathbf{m}_{11}^2, \ldots, \mathbf{m}_{1k}^2)$. With $2k$ rows, we obtain the first $2k$ moments of the k-sparse distribution corresponding to $\mathbf{m}_1$ (i.e., the real-valued, k-sparse distribution which places atomic probability π_j at $\mathbf{m}_{1j} \in \mathbb{R}$, also called a "$k$-spike" distribution). From this point on, one may apply the time-honored method of Prony to identify the source. For details, a runtime analysis, and further references, see Gordon et al. (2020).

Much of the interest of the present problem, by contrast, is due precisely to the fact that we *cannot* read out higher moments of the distributions corresponding to any of the bits X_i, because no relationship is assumed among the various rows of $\mathbf{m}$. How to nonetheless obtain higher moments of individual rows, is the challenge our algorithm tackles.

Main theorem In what follows ζ is an assumed separation parameter, and $\pi_{\min}$ is an assumed lower bound on mixture weights.

Theorem 1

(i) Given access to the joint statistics of n observable bits among which at least $3k-3$ which are ζ-separated, with all statistics available to additive accuracy ε for $\varepsilon \leq (\pi_{\min})^{O(\log k)}\zeta^{O(k^2 \log k)}$, our algorithm runs in time $2^{O(k^2)}n^{O(k)}$ and computes the model parameters (π and all row values $\mathbf{m}_{ij}$) to within accuracy $\varepsilon\zeta^{-O(k^2\log k)}(\pi_{\min})^{-O(\log k)}$.

(ii) If all rows are ζ-separated, then a slightly simpler version of our algorithm identifies the source (to within the same accuracy, given the same input accuracy, as in (i)), in runtime $2^{O(k^2)}n$.

(iii) If only $2k-1$ ζ-separated rows are available, another simpler version of our algorithm also identifies the source, but the loss factor on the accuracy is $(\pi_{\min})^{-k}\zeta^{-O(k^3)}$ and consequently one must start with $\varepsilon \leq (\pi_{\min})^{k}\zeta^{O(k^3)}$ which requires sample complexity comparable to prior work, but achieving the same improved runtime as in (i).

3. Algorithm

Further definitions It will be very useful to put our observables in matrix form. Let $S, T \subseteq [n]$ be disjoint sets and take any $\mathcal{A} \subseteq 2^S, \mathcal{B} \subseteq 2^T$. Then the matrix $\mathbf{C}_{\mathcal{B}\mathcal{A}}$ is observable (meaning every entry of it is a function of the joint statistics of the observable random variables $X_1, \ldots, X_n$), where

$$\mathbf{C}_{\mathcal{B}\mathcal{A}} \coloneqq \mathbf{M}[\mathcal{B}]\pi_{\odot}\mathbf{M}[\mathcal{A}]^{\top}. \tag{4}$$

Let $\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}} \coloneqq [\tilde{\mathrm{g}}(B \cup A)]_{B\in\mathcal{B}, A\in\mathcal{A}}$ be the corresponding matrix of empirical moments.

We use $\|\cdot\|$ to denote operator norm for matrices (w.r.t. Euclidean norm in the domain and range).

Strategy If $\mathbf{m}$ possesses $2k-1$ or $3k-3$ (depending on the version of the algorithm; this affects only runtime) ζ-separated rows of $\mathbf{m}$, and given empirical statistics within the required distance from perfect statistics, our algorithm will identify the model (to within similar accuracy). This includes even rows which are not themselves ζ-separated; all we need is that some $2k-1$ or $3k-3$ rows be ζ-separated. The algorithm has the following structure: range over all $n^{O(k)}$ subsets of the rows; run the identification algorithm using that set. If the set does not suffice for identification (which can happen only if the set includes some non-ζ-separated rows), this will be flagged by the algorithm. Any two such runs which do terminate successfully, must result in very close parameter reconstructions.

3.1. Synthetic bits and bootstrapping

We start with the $2k-1$-observables version of the algorithm. Our final algorithm in section 3.1.2, which uses $3k-3$ observables and achieves better runtime, will use a slightly more complicated construction, but the main ideas are present in the simpler variant here.

3.1.1. CONSTRUCTING HIGHER MOMENTS OF A ROW, USING $2k-2$ OTHER ROWS

In what follows we show how to compute moments of arbitrary degree of the k-spike distribution associated with any bit X_i, given access to any additional $2k-2$ ζ-separated observable bits. For concreteness, let X_1 (corresponding to the row $\mathbf{m}_1$) be the variable for which we want to find higher moments. (We don't require that $\mathbf{m}_1$ be ζ-separated for the moment computation, although that will be needed in a subsequent step of the algorithm.) Let $S = \{2, \ldots, k\}$, $T = \{k+1, \ldots, 2k-1\}$ be the indices of the $2k-2$ other ζ-separated rows, partitioned into two sets of $k-1$ rows each.

The only thing that our statistics tell us about row $\mathbf{m}_1$ in isolation is its first moment: $\mathbb{E}(X_1) = \mathbf{m}_1 \pi_\odot \mathbb{1}^\top$. Equivalently this quantity is also the expectation the k-spike distribution associated with X_1. It will be critical to obtain higher moments of this distribution. The second moment is equal to $(\mathbf{m}_1 \odot \mathbf{m}_1)\pi_\odot \mathbb{1}^\top$, and more generally (with $\mathbf{m}_1^{\odot r}$ denoting the r-fold Hadamard product of $\mathbf{m}_1$ with itself), the r'th moment is $\mathbf{m}_1^{\odot r}\pi_\odot \mathbb{1}^\top$ for any r. We will need to have the $1, \ldots, 2k$'th moments in order to solve for $\mathbf{m}_1$ and π.

We show in Appendix B that there exist subsets $\mathcal{A} = \{A_1, \ldots, A_k\} \subseteq 2^S$, $\mathcal{B} = \{B_1, \ldots, B_k\} \subseteq 2^T$ of size k each such that $\mathbf{A} \coloneqq \mathbf{M}[\mathcal{A}] \in \mathbb{R}^{k\times k}$ and $\mathbf{B} \coloneqq \mathbf{M}[\mathcal{B}] \in \mathbb{R}^{k\times k}$ are invertible. Moreover, we'll have $A_1 = \emptyset = B_1$ so that the first row of $\mathbf{M}[\mathcal{A}]$ and $\mathbf{M}[\mathcal{B}]$ will be the vector $\mathbb{1}$. Then the matrix $\mathbf{C}_{\mathcal{B}\mathcal{A}} = \mathbf{B}\pi_\odot \mathbf{A}^\mathsf{T}$ (as defined in (4)) is an invertible, observable matrix.

Now consider the vector

$$v_1 \coloneqq \mathbf{m}_1 \pi_\odot \mathbf{A}^\mathsf{T} = (\mathbb{E}[X_1 X_{A_1}], \ldots, \mathbb{E}[X_1 X_{A_k}]).$$

Each coordinate $\mathbb{E}[X_1 X_{A_i}] = \mathbb{E}[X_{A_i \cup \{1\}}]$ is a multi-linear moment and is therefore observable. In the algorithm, our starting point will be approximations to multi-linear moments, and we will address that shortly, but for now we write down the algorithm as it would run with perfect statistics.

While we'd like to find $\mathbf{m}_1^{\odot 2}$ directly, we won't be able to do so, and instead we'll settle for computing the vector u_1 satisfying

$$u_1 \mathbf{B} \coloneqq \mathbf{m}_1.$$

This u_1 is unique since $\mathbf{B}$ is invertible, and it gives the coefficients needed to express $\mathbf{m}_1$ as a linear combination of the rows of $\mathbf{B}$. We can in fact compute u_1 from observables, because

$$u_1 = v_1 \mathbf{C}_{\mathcal{B}\mathcal{A}}^{-1}. \tag{5}$$

We can verify (5) algebraically:

$$v_1(\mathbf{C}_{\mathcal{B}\mathcal{A}})^{-1}\mathbf{B} = v_1(A^\mathsf{T})^{-1}\pi_\odot^{-1}\mathbf{B}^{-1}\mathbf{B} = v_1(\mathbf{A}^\mathsf{T})^{-1}\pi_\odot^{-1} = \mathbf{m}_1\pi_\odot\mathbf{A}^\mathsf{T}(\mathbf{A}^\mathsf{T})^{-1}\pi_\odot^{-1} = \mathbf{m}_1. \tag{6}$$

Conceptually what (6) means is that $u_1\mathbf{B}$, a linear combination of rows of $\mathbf{B}$, defines a random variable that has been synthesized out of the X_{B_ℓ} random variables, and shares the same expectations as X_1 conditional on any setting of H. We call this the method of *synthetic bits*. Our algorithm consists of repeatedly *bootstrapping* synthetic bits.

To be more explicit, use $u_1 = (u_{11}, \ldots, u_{1k})$ to define a new random variable $Y \coloneqq \sum_{\ell=1}^k u_{1\ell} X_{B_\ell}$.

Y has been *synthesized* out of $X_{B_1}, \ldots, X_{B_k}$. Y has the same expectation as X_1, conditioned on the value of the hidden variable H, since

$$\mathbb{E}(X_1 \mid H = j) = (\mathbf{m}_1)_j = \sum_{\ell=1}^k u_{1\ell}\, \mathbb{E}(X_{B_\ell} \mid H = j) = \mathbb{E}(Y \mid H = j).$$

Moreover, given the value of H, X_1 and Y are independent, because $\{1\} \cap B_\ell = \emptyset$.

Bootstrapping to obtain the second moment. As a consequence of (6), we can perform the bootstrapping step which is at the heart of our algorithm. We are interested in obtaining the vector

$$v_2 \coloneqq \mathbf{m}_1^{\odot 2}\pi_{\odot}\mathbf{A}^{\mathsf{T}}$$

which is a linear image (under the mapping $\pi_{\odot}\mathbf{A}^{\mathsf{T}}$) of a random variable distributed as the product of two random variables which are independent conditional on H; each distributed as X_1 conditional on H. This we get by setting

$$v_2 \coloneqq (\mathbb{E}(X_1 Y X_{A_1}), \ldots, \mathbb{E}(X_1 Y X_{A_k})) \tag{7}$$

Since $A_1 = \emptyset$, the entry $(v_2)_1 = \mathbb{E}(X_1 Y) = \mathbf{m}_1^{\odot 2}\pi_{\odot}\mathbb{1}^{\mathsf{T}}$ is exactly our desired second moment. To get access to v_2, we observe that even though we don't necessarily have two independent copies of $\mathbf{m}_1$ among our rows, our synthetic bit Y provides the needed independence from X_1. As a matter of notation, for our collection $\mathcal{B} \subseteq 2^T$, we define $\mathcal{B} + \{1\}$ to consist of the sets $B_\ell \cup \{1\}$ for each $B_\ell \in \mathcal{B}$.

Now we can write

$$v_2 = \mathbf{m}_1^{\odot 2}\pi_{\odot}\mathbf{A}^{\mathsf{T}} = (\mathbf{m}_1 \odot (u_1\mathbf{B}))\pi_{\odot}\mathbf{A}^{\mathsf{T}} = u_1\mathbf{C}_{\mathcal{B}+\{1\},\mathcal{A}} \tag{8}$$

Since we already have u_1, and since $\mathbf{C}_{\mathcal{B}+\{1\},\mathcal{A}}$ is observable, expression (8) can be used to compute v_2.

Bootstrapping to all moments. The generalization of the second moment computation is this. Given vector u_{r-1} defined by $u_{r-1} \coloneqq \mathbf{m}_1^{\odot(r-1)}\mathbf{B}^{-1}$, we define v_r and u_r and show how to compute them:

Definition		Computation	
$v_r \coloneqq \mathbf{m}_1^{\odot r}\pi_{\odot}\mathbf{A}^{\mathsf{T}}$	$= (\mathbf{m}_1 \odot (u_{r-1}\mathbf{B}))\pi_{\odot}\mathbf{A}^{\top}$	$= u_{r-1}\mathbf{C}_{\mathcal{B}+\{1\},\mathcal{A}}$	(9)
$u_r \coloneqq \mathbf{m}_1^{\odot r}\mathbf{B}^{-1}$	$= \mathbf{m}_1^{\odot r}\pi_{\odot}\mathbf{A}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}-1}\pi_{\odot}^{-1}\mathbf{B}^{-1}$	$= v_r\mathbf{C}_{\mathcal{B}\mathcal{A}}^{-1}$.	(10)

In the actual algorithm, we'll be working with empirical approximations of v_r, u_r, and $\mathbf{C}_{\mathcal{B}\mathcal{A}}$, which we'll denote $\tilde{v}_r$, $\tilde{u}_r$, and $\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}$, respectively. A key part of the technical work will be in bounding error amplification. In order to compute $\tilde{v}_r$ and $\tilde{u}_r$, we'll use (following (9),(10)) the following assignments, where all quantities are empirical estimates:

$$\tilde{v}_r \coloneqq \tilde{u}_{r-1}\tilde{\mathbf{C}}_{\mathcal{B}+\{1\},\mathcal{A}} \tag{11}$$

$$\tilde{u}_r \coloneqq \tilde{v}_r(\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}})^{-1} \tag{12}$$

It is important to note that the bootstrapping can be performed only because $\tilde{u}_{r-1}$ places weight solely upon rows in $\mathcal{B} \subseteq 2^T$. Then the bootstrapping yields an expression for $\tilde{v}_r$ that contains moments involving X_1 as well as the entries in T. But once we compute $\tilde{u}_r$, we regain a coefficient vector for a synthetic bit, that again places weight solely upon rows in $\mathcal{B}$.

The fact that both the computation of $\tilde{v}_r$ (in (11)) and of $\tilde{u}_r$ (in (12)) work stably and not only in the perfect-statistics limit, relies upon the following:

Corollary 2 *If all empirical multilinear moments are within $\varepsilon < \zeta^{\Omega(k^2)}$ of their true values, then $\left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}} - \mathbf{C}_{\mathcal{B}\mathcal{A}}\right\| \leq k\varepsilon$ and $\left\|\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}^{-1} - \tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}^{-1}\right\| \leq \zeta^{-O(k^2)}\pi_{\min}^{-2}\varepsilon$.*

Proof This will be an immediate consequence of Lemma 11. ■

From this, we can bound the increase in error due to each subsequent application of (11) and (12).

Lemma 3 *If all empirical multilinear moments are within $\varepsilon < \zeta^{\Omega(k^2)}$ of their true values, then for all i,*

$$\|\tilde{u}_i - u_i\|_\infty < \zeta^{-O(k^2)}\pi_{\min}^{-2}\|\tilde{u}_{i-1} - u_{i-1}\|_\infty \quad \textit{and} \quad \|\tilde{v}_r - v_r\| \le \zeta^{-O(k^2)}\pi_{\min}^{-2}\|\tilde{v}_{i-1} - v_{i-1}\|_\infty .$$

Proof This follows from Lemma 12, which is itself a consequence of Lemma 11. ■

Nevertheless, if we compute each $\tilde{v}_r$ and $\tilde{u}_r$ as described above, we would need to start with accuracies $\varepsilon < \zeta^{\Omega(k^3)}$ in order to retain accuracy after bootstrapping $O(k)$ times, as is required by the above algorithm. To avoid this, we'll need to reduce the number of iterations by using $3k-3$, rather than $2k-1$, ζ-separated rows.

3.1.2. IMPROVED ERROR CONTROL: CONSTRUCTING HIGHER MOMENTS USING $3k-3$ ζ-SEPARATED ROWS

In order to avoid performing k iterations to compute $\tilde{v}_k$ and $\tilde{u}_k$, we'll use another set of rows, $\mathcal{B}' = \{B'_1, \ldots, B'_k\} \subseteq 2^{T'}$ of size k, where T' is disjoint from S and T, and $B'_1 = \emptyset$. Now we'll introduce $\mathbf{B}' \coloneqq \mathbf{M}[\mathbf{B}']$ and $\mathbf{C}_{\mathcal{B}'\mathcal{A}} \coloneqq \mathbf{B}'\pi_\odot\mathbf{A}^\mathsf{T}$, both of which will be invertible as before.

Previously, u_i was a linear combination of the rows of $\mathbf{B}$ such that $u_i\mathbf{B} = \mathbf{m}_1^{\odot i}$. We'll introduce a new, but similar, sequence of vectors u'_i where $u'_i\mathbf{B}' = \mathbf{m}_1^{\odot i}$ and u'_i is obtained from v_i by $u'_i = v_i\mathbf{C}_{\mathcal{B}'\mathcal{A}}^{-1}$. In the algorithm, we'll only have access to the approximations $\tilde{u}'_i$ and $\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}$ and we'll compute each successive $\tilde{u}'_i$ by $\tilde{u}'_i = \tilde{v}_i\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}^{-1}$.

The advantage we obtain over the more straightforward process of the previous section, results from our ability to compute v_{2^i} using $u_{2^{i-1}}$ and $u'_{2^{i-1}}$; in that way, we are able to get away with performing only $1 + \lg k$ iterations to compute any of $v_1, \ldots, v_{2k}$.

To describe the computation, we first define the sum of two collections of subsets $\mathcal{U}, \mathcal{V} \subseteq 2^{[n]}$ by

$$\mathcal{U} + \mathcal{V} \coloneqq \{U \cup V : U \in \mathcal{U}, V \in \mathcal{V}\} .$$

Now let x and y be vectors indexed by the the subsets in $\mathcal{B}$ and $\mathcal{B}'$, respectively. Recall that $\mathcal{B} \subseteq 2^T, \mathcal{B}' \subseteq 2^{T'}$, where $T \cap T' = \emptyset$. Then $|\mathcal{B} + \mathcal{B}'| = k^2$ and each subset in $\mathcal{B} + \mathcal{B}'$ can be uniquely written as $B_\ell \cup B'_j$ for $\ell, j \in [k]$. We define the Kronecker product $(x \otimes y) \in \mathbb{R}^{\mathcal{B}+\mathcal{B}'}$ to be the vector indexed by subsets in $\mathcal{B} + \mathcal{B}'$ given by

$$(x \otimes y)_{B_\ell \cup B'_j} \coloneqq x_{B_\ell} y_{B'_j}$$

for any $\ell, j \in [k]$. To access v_{2^i}, we write

$$v_{2^i} = \mathbf{m}_1^{\odot 2^i}\pi_\odot\mathbf{A}^\mathsf{T} = ((u_{2^{i-1}}\mathbf{B}) \odot (u'_{2^{i-1}}\mathbf{B}))\pi_\odot\mathbf{A}^\mathsf{T} = (u_{2^{i-1}} \otimes u'_{2^{i-1}})\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}},$$

expressing the row v_{2^i} as the linear of combination of the k^2 rows corresponding to the subsets in $\mathcal{B} + \mathcal{B}'$.

Of course, we'll need to compute v_ℓ for ℓ not a power of 2. We can do this using a slight modification of the recursive procedure where for $i = 1, \ldots, 1 + \lg k$ and $j = 1, \ldots, 2^i$, we compute

$$\tilde{v}_\ell \coloneqq (\tilde{u}_j \otimes \tilde{u}'_{2^i})\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} \tag{13}$$

where $\ell = j + 2^i$, and we've computed $\tilde{u}_j$ in a prior iteration for all $j \le 2^i$.

Under this modification, each $\tilde{v}_i$ is produced in only $1 + \lg k$ iterations each of which involves a matrix multiplication by $\tilde{\mathbf{C}}^{-1}_{\mathcal{BA}}$ or a convolution followed by multiplication by $\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$, each step of which can increase the error in total by $\zeta^{-O(k^2 \log k)}\pi_{\min}^{-2}$. By starting with empirical moments accurate to within $\zeta^{\Omega(k^2 \log k)}\pi_{\min}^{2k}$, we can ensure that the resulting vectors $\tilde{v}_i$ and $\tilde{u}_i$ are sufficiently close to start solving for $\mathbf{m}_1$ and π.

We provide pseudocode for this "$3k-3$ rows" version of the algorithm, see Fig. 1.

Flagging a failure condition If the chosen rows $S \cup T \cup T'$ fail to all be ζ-separated, the algorithm might fail. However, we will detect such failure. The conditions that we actually need so that the algorithm should work, are these: (a) $\tilde{\mathbf{C}}_{\mathcal{BA}}$ and $\tilde{\mathbf{C}}_{\mathcal{B}',\mathcal{A}}$ should have a large least singular value. (It does not actually matter whether all rows we use are ζ-separated, that was merely a sufficient condition for this well-conditioning.) We compute this singular value explicitly and simply dismiss the triple S, T, T' if this condition fails. (b) The first row of $\mathcal{B}$, namely $\{1\}$ in the numbering used in the pseudocode for Algorithm 1, should be ζ-separated. If condition (a) holds but this condition fails, we will detect the failure in line 12 of the algorithm, because the Hankel matrix will have insufficient eigenvalue gap (see Cor. 12 of Gordon et al. (2020)).

3.2. Solving the power distribution problem

Once we've computed $\tilde{v}_1, \ldots, \tilde{v}_{2k}$, we have access to all of the moments of the k-spike distribution corresponding to observable X_1. (The ith moment equals $(\tilde{v}_i)_1$, just as in Eqn. (7)). Recall, these are the moments of a mixture of k Bernoulli random variables, where the r'th moment corresponds to drawing a mixture component j with probability π_j, then setting the Bernoulli random variable to 1 with probability $\mathbf{m}_{1j}^r$. The problem of recovering the parameters (i.e., the vectors $\mathbf{m}_1$ and π) from approximate moments of this form has been extensively studied, and many algorithms have been provided. We use the algorithm LEARNPOWERDISTRIBUTION from Gordon et al. (2020), which on inputs accurate to within ε, outputs parameters $\tilde{\mathbf{m}}_1$ and $\tilde{\pi}$ to within accuracy $\frac{1}{\pi_{\min}}\zeta^{-O(k)}\varepsilon$ while running in time (arithmetic operations) $k^{2+o(1)}$.

3.3. Recovering the remaining parameters

Once we have estimates for $\mathbf{m}_1$ and π, we can simply solve for $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{B}'$ using the fact that $\mathbf{A}^\mathsf{T} = \pi_\odot^{-1}(\mathrm{Vdm}(\mathbf{m}_1))^{-1}V$ where

$$V = (v_0; \ldots; v_{k-1})$$

is the matrix with rows v_i and $\mathrm{Vdm}(\mathbf{m}_1)$ is the Vandermonde matrix with rows $\mathbf{m}_1^{\odot i}$ for $i = 0, \ldots, k-1$. Note that here we finally do require that $\mathbf{m}_1$ be ζ-separated. We can thus solve for $\mathbf{A}^\mathsf{T}$.

To solve for $\mathbf{B}$ we use $\mathbf{B} = \mathbf{C}_{\mathcal{BA}}(\mathbf{A}^\mathsf{T})^{-1}\pi_\odot^{-1}$. Likewise for $\mathbf{B}'$ we use $\mathbf{B}' = \mathbf{C}_{\mathcal{B}'\mathcal{A}}(\mathbf{A}^\mathsf{T})^{-1}\pi_\odot^{-1}$.

Now for any row i not already computed, we need only pick any other set $\mathcal{S} = \{S_1, \ldots, S_k\}$ of k linearly independent rows supported on a set not containing i, and we can solve for $\mathbf{m}_i$ by writing

$$\mathbf{m}_i = (\mathbb{E}[X_i X_{S_1}], \ldots, \mathbb{E}[X_i X_{S_k}])\mathbf{M}[\mathcal{S}]^{\mathsf{T}^{-1}}\pi_\odot^{-1}$$

In particular, by setting $\mathcal{S} = \mathcal{A}$ we can solve for all rows in $[n] \setminus S$ and by setting $\mathcal{S} = \mathcal{B}$ we can solve for all rows in $[n] \setminus T$. Together, this suffices to solve for all rows.

3.4. Runtime

The algorithm contains three main parts:

1. Find disjoint $S, T, T' \subset [n]$ and $\mathcal{A} \subset 2^S$, $\mathcal{B} \subset 2^T$, and $\mathcal{B}' \subset 2^{T'}$, which is complexity $n^{O(k)}2^{O(k^2)}$. First we require $n^{O(k)}$ iterations to check all possible disjoint S, T, T'. Then, $2^{O(k^2)}$ operations are required in each iteration to check all size k subsets of the 2^{k-1} rows of $\mathbf{M}[2^S], \mathbf{M}[2^T]$ and $\mathbf{M}[2^{T'}]$.

2. Nested loops to compute higher order moments $\tilde{v}$, and the corresponding $\tilde{u}$. This step takes time $O(\text{poly}(k))$.

3. Applying the power distribution result. This can be done in time $O(k^{2+o(1)}+k(\log^2 k)\log\log(\varepsilon^{-1}))$ (see Corollary 16).

This gives runtime complexity of $n^{O(k)}2^{O(k^2)} + O(k^{2+o(1)} + k(\log^2 k)\log\log(\varepsilon^{-1}))$. If all sources are ζ-separated, we do not need to iterate over choices of S, T, T', so the runtime improves to $2^{O(k^2)} + O(k^{2+o(1)} + k\log^2(k)\log\log(\varepsilon^{-1}))$.

3.5. Analyzing the algorithm

The analysis of the algorithm appears in Appendix A. Crucially, the analysis will depend on the condition number bound for the matrix $\mathbf{M}$, stated below and proved in Appendix B.

3.6. The core stability bound

Definition 4 *We define* $\beta := \frac{(\zeta/2)^{k-1}}{k}$. *We will often simplify expressions with the bound* $\beta \geq \zeta^{2k}$.

Theorem 5 *Let S be a set of $k-1$ ζ-separated vectors $\mathbf{m}_1, \ldots, \mathbf{m}_{k-1}$. Then there exists a set $J \subseteq 2^S$, $|J| = k$, such that $\sigma_k(\mathbf{M}[J]) \geq \beta^k 2^{-3k/2}k^{-3/2}$ and $\sigma_{\max}(\mathbf{M}[J]) \leq k$. The first row of $\mathbf{M}[J]$ is $\mathbb{1}$ (corresponding to $\emptyset \in J$).*

Bounding the condition number of $\mathbf{M}[2^S]$. Most of our work will go into lower bounding the k'th singular value of $\mathbf{M}[2^S]$. This is where the ζ-separation condition is essential. (Even the non-quantitative result that $\mathbf{M}[2^S]$ has full column rank is not trivial; indeed, the result of Tahmasebi et al. (2018) that sources can be identified from $2k-1$ separated bits, is implied by the non-quantitative version of this section, omitted here, which employs the same approach but is considerably shorter.)

The proof of Theorem 5 is in Appendix B.

Let $\mathcal{B} \subseteq 2^{\{1,\dots,k-1\}}, \mathcal{B}' \subseteq 2^{\{k,\dots,2k-2\}}, \mathcal{A} \subseteq 2^{\{2k-1,\dots,3k-3\}}$, with $|\mathcal{A}| = |\mathcal{B}| = |\mathcal{B}'| = k$ maximizing $\min\{\sigma_k(\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}), \sigma_k(\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}})\}$.
If this min is below $\pi_{\min}\zeta^{O(k^2)}$, terminate.
Denote $\mathcal{B} = \{B_1, \dots, B_k\}$, $\mathcal{B}' = \{B'_1, \dots, B'_k\}$, and $\mathcal{A} = \{A_1, \dots, A_k\}$. Without loss of generality $B_1 = \{1\}$.
$\tilde{v}_0 \leftarrow (\tilde{\mathrm{g}}(A_1); \dots ; \tilde{\mathrm{g}}(A_k))$.
$\tilde{v}_1 \leftarrow (\tilde{\mathrm{g}}(A_1 \cup \{1\}); \dots ; \tilde{\mathrm{g}}(A_k \cup \{1\}))$.
$\tilde{u}_1 \leftarrow \tilde{v}_1(\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}})^{-1}$.
$\tilde{u}'_1 \leftarrow \tilde{v}_1(\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}})^{-1}$.
for $i = 1, \dots, 1 + \lg k$ **do**
  for $j = 1, \dots, 2^{i-1}$ **do**
    $\tilde{v}_{2^{i-1}+j} \leftarrow (\tilde{u}_j \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$.
    $\tilde{u}_{2^{i-1}+j} \leftarrow \tilde{v}_{2^{i-1}+j}(\tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}})^{-1}$.
  end
  $\tilde{u}'_{2^i} \leftarrow \tilde{v}_{2^i}(\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}})^{-1}$.
end
Let H_{k+1} be the $(k+1) \times (k+1)$ Hankel matrix with entries given by $[H_{k+1}]_{i,j=0}^{k} = (\tilde{v}_{i+j})_1$.
If the second-smallest eigenvalue of H_{k+1} is below $\frac{\pi_{\min}}{2}(\zeta/16)^{2k-2}$, terminate.
$\tilde{\mathbf{m}}_1, \tilde{\pi} \leftarrow$ LEARNPOWERDISTRIBUTION(H_{k+1}).
$\tilde{V} \leftarrow (\tilde{v}_0; \dots ; \tilde{v}_{k-1})$.
$\mathrm{Vdm}(\tilde{\mathbf{m}}_1) \leftarrow (\tilde{\mathbf{m}}_1^{\odot 0}; \dots ; \tilde{\mathbf{m}}_1^{\odot(k-1)})$.
$\tilde{\mathbf{A}}^{\mathsf{T}} \leftarrow \tilde{\pi}_{\odot}^{-1}(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\tilde{V}$.
$\tilde{\mathbf{B}} \leftarrow \tilde{\mathbf{C}}_{\mathcal{B}\mathcal{A}}(\tilde{\mathbf{A}}^{\mathsf{T}})^{-1}\pi_{\odot}^{-1}$.
$\tilde{\mathbf{B}}' \leftarrow \tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}(\tilde{\mathbf{A}}^{\mathsf{T}})^{-1}\pi_{\odot}^{-1}$.
For every $i \in [n] \setminus [k]$, $\tilde{\mathbf{m}}_i \leftarrow (\tilde{\mathrm{g}}(A_1 \cup \{i\}), \dots, \tilde{\mathrm{g}}(A_k \cup \{i\}))^{\mathsf{T}} \left(\tilde{\mathbf{A}}^{\mathsf{T}}\right)^{-1} \tilde{\pi}_{\odot}^{-1}$.
For every $i \in \{2, \dots, k\}$, $\tilde{\mathbf{m}}_i \leftarrow (\tilde{\mathrm{g}}(B_1 \cup \{i\}), \dots, \tilde{\mathrm{g}}(B_k \cup \{i\}))^{\mathsf{T}} \left(\tilde{\mathbf{B}}^{\mathsf{T}}\right)^{-1} \tilde{\pi}_{\odot}^{-1}$.

Algorithm 1: Identifies a mixture of product distributions given $3k-3$ ζ-separated observable bits

Acknowledgments

Research supported in part by NSFC-ISF grant 2553-17, NSF-BSF grant 2018687, and NSF grants CCF-1618795 and 1909972. Part of this work was done while the third author visited Caltech. Thanks to anonymous reviewers for helpful comments.

References

E. Allman, C. Matias, and J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37:3099–3132, 2009.

A. Anandkumar, D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu. A spectral algorithm for latent Dirichlet allocation. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 917–925. Curran Associates, Inc., 2012a.

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, 2012b.

B. Aragam, C. Dan, E. P. Xing, and P. Ravikumar. Identifiability of nonparametric mixture models and Bayes optimal clustering. *Ann. Statist.*, 48(4):2277–2302, 2020.

S. Arora, R. Ge, and A. Moitra. Learning topic models — going beyond SVD. In *Proc. 53rd Ann. IEEE Symp. on Foundations of Computer Science*, 2012.

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

M. A. Carreira-Perpiñán and S. Renals. Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12(1):141–152, 2000.

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.

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.

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.

S. Dasgupta. Learning mixtures of Gaussians. In *Proc. 40th Ann. IEEE Symp. on Foundations of Computer Science*, page 634–644, 1999.

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

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.

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.

S. L. Gordon, B. Mazaheri, Y. Rabani, and L. J. Schulman. The sparse Hausdorff moment problem, with application to topic models. ArXiv:2007.08101, 2020.

Y. Ji, C. Wu, P. Liu, J. Wang, and K. R. Coombes. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, 2005.

A. Juan and E. Vidal. On the use of Bernoulli mixture models for text classification. *Pattern Recognition*, 35(12):2705–2710, 2002.

A. Juan and E. Vidal. Bernoulli mixture models for binary images. In *Proc. of the 17th International Conference on Pattern Recognition*, volume 3, pages 367–370, 2004.

A. Juan, J. García-Hernández, and E. Vidal. EM initialisation for Bernoulli mixture learning. In A. Fred, T. M. Caelli, R. P. W. Duin, A. C. Campilho, and D. de Ridder, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, pages 635–643, Berlin, Heidelberg, 2004. Springer.

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.

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

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

T. C. Koopmans. *Statistical Inference in Dynamic Economic Models*. John Wiley and Sons, Inc., 1950.

T. C. Koopmans and O. Reiersol. The identification of structural characteristics. *Ann. Math. Statist.*, 21(2):165–181, 1950.

C. Li, B. Wang, V. Pavlu, and J. Aslam. Conditional Bernoulli mixtures for multi-label classification. In *Proc. of the 33rd International Conference on Machine Learning*, pages 2482–2491, 2016.

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.

B. G. Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

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.

A. Najafi, S. A. Motahari, and H. R. Rabiee. Reliable clustering of Bernoulli mixture models. *Bernoulli*, 26(2):1535–1559, 2020.

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.

J. A. Palmer, K. Kreutz-Delgado, and S. Makeig. A generalized multivariate logistic model and EM algorithm based on the normal variance mean mixture representation. In *IEEE Statistical Signal Processing Workshop*, pages 1–5, 2016.

J. Pearl. *Causality*. Cambridge, 2nd edition, 2009.

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

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

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.

A. Ritchie, R. A. Vandermeulen, and C. D. Scott. Consistent estimation of identifiable nonparametric mixture models from grouped observations. *CoRR*, abs/2006.07459, 2020.

K. Schmüdgen. *The Moment Problem*, volume 277 of *Graduate Texts in Mathematics*. Springer International Publishing, 2017.

B. Simon. *A comprehensive course in analysis*. American Mathematical Society, 2015.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, second edition, 2000.

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.

H. Teicher. Identifiability of finite mixtures. *Ann. Math. Statist.*, 34(4):1265–1269, 12 1963.

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

R. A. Vandermeulen and C. Scott. On the identifiability of mixture models from grouped samples. *ArXiv*, abs/1502.06644, 2015.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *Ann. Math. Statist.*, 39(1): 209–214, 02 1968.

Appendix A. Analyzing the Algorithm

As we have seen the algorithm consists of bootstrapping steps which, as indicated in Eqns. (11), (12), lift us "forward" from $\tilde{u}_{r-1}$ to $\tilde{v}_r$ and then "back" to $\tilde{u}_r$. We must now control the loss in accuracy of the statistics, in each of these steps. It turns out that the first of these is easier and less expensive in accuracy; while the second, in which we "invert" from approximate statistics to obtain a linear combination of sources, is harder and also more expensive. In this section we show how to control these steps. We rely for this control on a condition number bound (which is not in itself algorithmic and is due entirely to the ζ-separation), which will be given in Appendix B. Throughout the analysis, we'll assume that every multi-linear moment we use is known with additive error bounded by

$$\varepsilon := \zeta^{C_1 k^2 \log k} \pi_{\min}^{C_2 \log k}$$

for constants C_1, C_2. Choosing $C_1 := 60, C_2 := 8$ is sufficient to give us final error $\zeta^{\Omega(k^2 \log k)}$.

A.1. Bounding $\|\tilde{u}_j - u_j\|$ for $j \leq 2k$

The following Lemma is a consequence of Theorem 5 to be proven in the next section:

Lemma 6 *When the input mixture contains $3k-3$ ζ-separated rows, we can find disjoint sets $S, T, T' \subseteq [n]$ of size $k-1$ each and subsets $\mathcal{A} \subseteq 2^S, \mathcal{B} \subseteq 2^T, \mathcal{B}' \subseteq 2^{T'}$ with $|\mathcal{A}|, |\mathcal{B}| = k$ such that the matrices $\mathbf{A} := \mathbf{M}[\mathcal{A}]$, $\mathbf{B} := \mathbf{M}[\mathcal{B}]$, and $\mathbf{B}' := \mathbf{M}[\mathcal{B}']$ satisfy*

1. *The first row of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{B}'$ is the all-ones vector, $\mathbb{1}$.*
2. *$\sigma_k(\mathbf{A}), \sigma_k(\mathbf{B}), \sigma_k(\mathbf{B}') \geq \beta^k 2^{-3k/2} k^{-3/2}$.*
3. *$\sigma_{\max}(\mathbf{M}[\mathcal{A}]), \sigma_{\max}(\mathbf{M}[\mathcal{B}]), \sigma_{\max}(\mathbf{M}[\mathcal{B}]) \leq k$.*

And the matrices $\mathbf{C}_{\mathcal{B}\mathcal{A}} = \mathbf{B}\pi_{\odot}\mathbf{A}^{\mathsf{T}}$ and $\mathbf{C}_{\mathcal{B}'\mathcal{A}} = \mathbf{B}'\pi_{\odot}\mathbf{A}^{\mathsf{T}}$ satisfy

1. *$\sigma_{\max}(\mathbf{C}_{\mathcal{B}\mathcal{A}}), \sigma_{\max}(\mathbf{C}_{\mathcal{B}'\mathcal{A}}) \leq k^2$.*
2. *$\sigma_k(\mathbf{C}_{\mathcal{B}\mathcal{A}}), \sigma_k(\mathbf{C}_{\mathcal{B}'\mathcal{A}}) \geq \beta^{2k} 2^{-3k} k^{-3} \pi_{\min}$.*

Proof This follows immediately from Theorem 5, the definition $\mathbf{C}_{\mathcal{B}\mathcal{A}} = \mathbf{M}[\mathcal{B}]\pi_{\odot}\mathbf{M}[\mathcal{A}]^{\mathsf{T}}$, and the min-max characterization of the first and last singular values. ■

Corollary 7 $\left\|(\mathbf{C}_{\mathcal{B}\mathcal{A}})^{-1}\right\| \leq \zeta^{-10k^2}\pi_{\min}^{-1}$.

Proof $\left\|(\mathbf{C}_{\mathcal{B}\mathcal{A}})^{-1}\right\| \leq (\zeta^{-3k})^{2k} 2^{3k} k^3 \pi_{\min}^{-1} \leq \zeta^{-10k^2}\pi_{\min}^{-1}$ ■

Corollary 8 $\left\|\mathbf{A}^{-1}\right\|, \left\|\mathbf{B}^{-1}\right\|, \left\|(\mathbf{B}')^{-1}\right\| \leq \zeta^{-6k^2}$.

Lemma 9 $\|u_i\| \leq \zeta^{-6k^2}$, *and* $\|v_i\| \leq \zeta^{-1}$.

Proof We observe that $\|u_i\| = \left\|\mathbf{m}_1^{\odot i}\mathbf{B}^{-1}\right\| \leq k\left\|\mathbf{B}^{-1}\right\| \leq \beta^{-k}2^{3k/2}k^{3/2} \leq \zeta^{-6k^2}$. On the other hand, $\|v_i\| \leq k \leq \zeta^{-1}$, since v_i is a vector of moments of products of Bernoulli random variables. ■

Proposition 10 $\|\mathbf{C}_{\mathcal{BA}}\|, \left\|\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| \leq k^3$. *If all moments are within ε of their true values,* $\left\|\tilde{\mathbf{C}}_{\mathcal{BA}}\right\|, \left\|\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| \leq 2k^3$.

Lemma 11 *If all multilinear moments are within ε of their true values, then*

$$\left\|\tilde{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\|_2 \leq \zeta^{-1}\varepsilon, \qquad \left\|\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} - \mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| \leq \zeta^{-2}\varepsilon,$$

and

$$\left\|\tilde{\mathbf{C}}_{\mathcal{BA}}^{-1} - \mathbf{C}_{\mathcal{BA}}^{-1}\right\| \leq \zeta^{-26k^2}\pi_{\min}^{-2}\varepsilon.$$

Proof The first two inequalities just use $\|\cdot\|_2 \leq \|\cdot\|_F$. For the final inequality we use Lemma 33:

$$\left\|\tilde{\mathbf{C}}_{\mathcal{BA}}^{-1} - \mathbf{C}_{\mathcal{BA}}^{-1}\right\| \leq 2\left\|\mathbf{C}_{\mathcal{BA}}^{-1}\right\|^2 \left\|\mathbf{C}_{\mathcal{BA}} - \tilde{\mathbf{C}}_{\mathcal{BA}}\right\| \leq \beta^{-4k}2^{6k+1}k^7\pi_{\min}^{-2}\varepsilon$$

■

Lemma 12 *When the assumptions of Lemma 11 are satisfied, we have for any $i \in [2k]$ and $j = \lceil \log i \rceil$,*

$$\|\tilde{v}_i - v_i\|, \|\tilde{u}_j - u_j\|, \left\|\tilde{u}'_j - u'_j\right\| \leq \zeta^{-42ik^2}\pi_{\min}^{-2i}\varepsilon.$$

Proof Recall that we initialize the algorithm with

$$\tilde{v}_1 \leftarrow (\tilde{\mathrm{g}}(A_1 \cup \{1\}), \ldots, \tilde{\mathrm{g}}(A_k \cup \{1\})), \qquad \tilde{u}_1 \leftarrow \tilde{v}_1(\tilde{\mathbf{C}}_{\mathcal{BA}})^{-1}, \qquad \tilde{u}'_1 \leftarrow \tilde{v}_1(\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}})^{-1}.$$

First, we observe that $\|\tilde{v}_1 - v_1\| \leq \varepsilon$ by assumption. Since $\tilde{u}_1, \tilde{u}'_1$ are computed in the same manner here as in the loop, we'll bound that error in the induction. Now assume that the claim holds up to $i-1$. Recall that in each iteration of the outer loop we compute

$$\tilde{v}_{2^i} \leftarrow (\tilde{u}_{2^{i-1}} \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}, \qquad \tilde{u}_{2^i} \leftarrow \tilde{v}_{2^i}(\tilde{\mathbf{C}}_{\mathcal{BA}})^{-1}, \qquad \tilde{u}'_{2^i} \leftarrow \tilde{v}_{2^i}\tilde{\mathbf{C}}_{\mathcal{B}'\mathcal{A}}.$$

We'll first focus on bounding $\|\tilde{v}_{2^i} - v_{2^i}\|_\infty$. To do this we write

$$\tilde{v}_{2^i} - v_{2^i} = (\tilde{u}_{2^{i-1}} \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} - (u_{2^{i-1}} \otimes u'_{2^{i-1}})\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$$

and letting $w = \tilde{u}_{2^{i-1}} - u_{2^{i-1}}$, $w' = \tilde{u}'_{2^{i-1}} - u_{2^{i-1}}$, and $E = \tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} - \mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}$ we can bound the norm of the difference as follows, using the bilinearity of the Kronecker product, Lemma 9, and the induction hypothesis:

$$\begin{aligned}
\|\tilde{v}_{2^i} - v_{2^i}\| &= \left\|(\tilde{u}_{2^{i-1}} \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}} - (u_{2^{i-1}} \otimes u'_{2^{i-1}})\mathbf{C}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| \\
&\leq \left\|(w \otimes \tilde{u}'_{2^{i-1}})\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| + \left\|(\tilde{u}_{2^{i-1}} \otimes w')\tilde{\mathbf{C}}_{\mathcal{B}+\mathcal{B}',\mathcal{A}}\right\| + \left\|(\tilde{u}_{2^{i-1}} \otimes \tilde{u}'_{2^{i-1}})E\right\| \\
&\leq 2\zeta^{-42(i-1)k^2}\pi_{\min}^{-2(i-1)}\zeta^{-6k^2}k^3\varepsilon + \zeta^{-12k^2}\zeta^{-2}\varepsilon \\
&\leq \zeta^{-(42(i-1)+16)k^2}\pi_{\min}^{-2(i-1)}\varepsilon
\end{aligned}$$

Now we can bound $\|\tilde{u}_{2^i} - u_{2^i}\|_\infty$ by observing that

$$\tilde{u}_{2^i} - u_{2^i} = \tilde{v}_{2^i}\tilde{\mathbf{C}}_{\mathcal{BA}}^{-1} - v_{2^i}\mathbf{C}_{\mathcal{BA}}^{-1}.$$

Let $z = \tilde{v}_{2^i} - v_{2^i}$ and let $D = \tilde{\mathbf{C}}_{\mathcal{BA}}^{-1} - \mathbf{C}_{\mathcal{BA}}^{-1}$. The above equation becomes

$$\tilde{u}_{2^i} - u_{2^i} = (v_{2^i} + z)(\mathbf{C}_{\mathcal{BA}}^{-1} + D) - v_{2^i}\mathbf{C}_{\mathcal{BA}}^{-1} = v_{2^i}\mathbf{C}_{\mathcal{BA}}^{-1} + z\mathbf{C}_{\mathcal{BA}}^{-1} + zD$$

and after taking norms and using the triangle inequality we obtain

$$\|\tilde{u}_{2^i} - u_{2^i}\| \le \|v_{2^i}D\| + \left\|z\mathbf{C}_{\mathcal{BA}}^{-1}\right\| + \|zD\|$$

By Corollary 7, Lemma 9 and the induction hypothesis, we get

$$\begin{aligned}
\|\tilde{u}_{2^i} - u_{2^i}\| &\le \zeta^{-1}\zeta^{-26k^2}\pi_{\min}^{-2}\varepsilon + \zeta^{-(42(i-1)+16)k^2}\pi_{\min}^{-2(i-1)}\varepsilon\zeta^{-16k^2}\pi_{\min}^{-1} \\
&\quad + \zeta^{-(42(i-1)+16)k^2}\pi_{\min}^{-2(i-1)}\zeta^{-26k^2}\pi_{\min}^{-2}\varepsilon \\
&\le \zeta^{-26k^2-1}\pi_{\min}^{-2}\varepsilon + \zeta^{-(42(i-1)+20)k^2}\pi_{\min}^{-2(i-1)-1}\varepsilon + \zeta^{-(42(i-1)+42)k^2}\pi_{\min}^{-2(i-1)}\pi_{\min}^{-2}\varepsilon \\
&\le \zeta^{-42ik^2}\pi_{\min}^{-2i}\varepsilon
\end{aligned}$$

For j not a power of 2, we can do the same analysis, and since the error bound is increasing in j, the result will follow. ■

Corollary 13 *Algorithm 1 will produce vectors $\tilde{v}_i$ for $i \le 2k$ satisfying*

$$\|\tilde{v}_i - v_i\| \le \zeta^{-42k^2(1+\lg k)}\pi_{\min}^{-2(1+\lg k)}\varepsilon.$$

A.2. Applying the power distribution result

Definition 14 *Given a mixture $\mathcal{M}$ of k Bernoulli random variables with probabilities $m_1, \ldots, m_k$ and mixing probabilities $\pi_1, \ldots, \pi_k$, respectively, let $[\mathcal{H}_{k+1}]_{i,j=0}^{k} = \mu_{i+j}$ be the matrix of moments of the distribution.*

Theorem 15 (Theorem 17 from Gordon et al. (2020)) *Given a mixture $\mathcal{M} = (m, \pi)$ as above where m is ζ-separated, there is an algorithm,* LEARNPOWERDISTRIBUTION*, that takes a Hankel matrix $[\tilde{\mathcal{H}}_{k+1}]_{i,j=0}^{k} = \tilde{\mu}_{i+j}$ of approximate moments of $\mathcal{M}$ satisfying $\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\|_2 \le \pi_{\min}2^{-\gamma}\zeta^{16k}$ (for some $\gamma \ge 1$) and outputs a model $\tilde{\mathcal{M}} = (\tilde{m}, \tilde{\pi})$ satisfying*

$$\|\tilde{m} - m\|_\infty, \|\tilde{\pi} - \pi\|_\infty \le 2^{-\gamma}$$

using $O(k^2 \log k + k \log^2 k \cdot \log(\log \zeta^{-1} + \log \pi_{\min}^{-1} + \gamma))$ arithmetic operations.

Corollary 16 *The output $(\tilde{\mathbf{m}}_1, \tilde{\pi})$ of* LEARNPOWERDISTRIBUTION *in line 14 of Algorithm 1 will satisfy*

$$\|\tilde{\mathbf{m}}_1 - \mathbf{m}_1\|, \|\tilde{\pi} - \pi\| \le \zeta^{-42k^2(1+\lg k)-16k-1}\pi_{\min}^{-3-2\lg k}\varepsilon.$$

This step will use $O(k^2 \log k + k \log^2 k \cdot \log\log(\varepsilon^{-1}))$ arithmetic operations.

Proof Every entry $(\tilde{v}_i)_1$ satisfies $\|(\tilde{v}_i)_1 - (v_i)_1\| \le \zeta^{-42k^2(1+\lg k)}\pi_{\min}^{-2(1+\lg k)}\varepsilon$ so

$$\left\|\tilde{\mathcal{H}}_{k+1} - \mathcal{H}_{k+1}\right\| \le \zeta^{-42k^2(1+\lg k)}\pi_{\min}^{-2(1+\lg k)}\varepsilon$$

which implies that

$$\|\tilde{\mathbf{m}}_1 - \mathbf{m}_1\|_\infty, \|\tilde{\pi} - \pi\|_\infty \le \zeta^{-42k^2(1+\lg k)-16k}\pi_{\min}^{-2(1+\lg k)-1}\varepsilon.$$

Finally, we add a factor of ζ^{-1} to convert back to the Euclidean norm to get the stated bound. ■

A.3. Solving for the rest of the model

Once we've computed $\tilde{\mathbf{m}}_1$ and $\tilde{\pi}$, we'll use them to compute the remaining model parameters. In this section we bound the additional error introduced by these computations.

Proposition 17 $\|\mathrm{Vdm}(\tilde{\mathbf{m}}_1) - \mathrm{Vdm}(\mathbf{m}_1)\| \le \zeta^{-1}\|\tilde{\mathbf{m}}_1 - \mathbf{m}_1\|$.

Claim 18 (Claim 26 in Gordon et al. (2020)) $\left\|\mathrm{Vdm}(\mathbf{m}_1)^{-1}\right\| \le 2^k/\zeta^{k-1} \le \zeta^{-2k}$ *when* $\mathbf{m}_1$ *is* ζ*-separated.*

Lemma 19 *The computed* $\tilde{\mathbf{A}}$ *produced by Algorithm 1 will satisfy*

$$\left\|\tilde{\mathbf{A}} - \mathbf{A}\right\| \le \zeta^{-42k^2(1+\lg k)-20k-6}\pi_{\min}^{-5-2\lg k}\varepsilon.$$

Proof Recall $\tilde{V} = (\tilde{v}_0; \ldots; \tilde{v}_{k-1})$ from Algorithm 1 and $V = (v_0; \ldots; v_{k-1})$ is its real-value analog. First, we observe that $\left\|\tilde{V}\right\| \le \|V\| \le \zeta^{-2}$ and $\left\|\mathrm{Vdm}(\tilde{\mathbf{m}}_1)^{-1}\right\| \le \zeta^{-3k}$ by Lemma 33 and Claim 18. Now $\left\|\tilde{\pi}_\odot^{-1} - \pi_\odot^{-1}\right\| \le \zeta^{-42k^2(1+\lg k)-16k-2}\pi_{\min}^{-5-2\lg k}\varepsilon$ by Lemma 33 and Corollary 16. Thus,

$$\left\|\tilde{\pi}_\odot^{-1} - \pi_\odot^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V}\right\| \le \zeta^{-42k^2(1+\lg k)-18k-5}\pi_{\min}^{-5-2\lg k}\varepsilon.$$

Now $\left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1} - (\mathrm{Vdm}(\mathbf{m}_1))^{-1}\right\| \le \zeta^{-42k^2(1+\lg k)-20k-2}\pi_{\min}^{-3-2\lg k}\varepsilon$ by Lemma 33, so

$$\left\|\tilde{\pi}_\odot^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1} - (\mathrm{Vdm}(\mathbf{m}_1))^{-1}\right\| \left\|\tilde{V}\right\| \le \zeta^{-42k^2(1+\lg k)-20k-4}\pi_{\min}^{-4-2\lg k}\varepsilon.$$

Finally, $\left\|\tilde{V} - V\right\| \le \zeta^{-42k^2(1+\lg k)-1}\pi_{\min}^{1-2\lg k}\varepsilon$ so that

$$\left\|\pi_\odot^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V} - V\right\| \le \zeta^{-42k^2(1+\lg k)-3k-1}\pi_{\min}^{-3-2\lg k}\varepsilon.$$

Putting these together, we easily obtain

$$\begin{aligned}
\left\|\tilde{\mathbf{A}} - \mathbf{A}\right\| &= \left\|\tilde{\pi}_\odot^{-1}(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\tilde{V} - \pi_\odot^{-1}(\mathrm{Vdm}(\mathbf{m}_1))^{-1}V\right\| \\
&\le \left\|\tilde{\pi}_\odot^{-1} - \pi_\odot^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V}\right\| + \left\|\tilde{\pi}_\odot^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1} - (\mathrm{Vdm}(\mathbf{m}_1))^{-1}\right\| \left\|\tilde{V}\right\| \\
&\quad + \left\|\pi_\odot^{-1}\right\| \left\|(\mathrm{Vdm}(\tilde{\mathbf{m}}_1))^{-1}\right\| \left\|\tilde{V} - V\right\| \left(\left\|\mathrm{Vdm}(\mathbf{m}_1)^{-1}E_2\right\| + \|E_1 V\|_\infty\right) \\
&\quad + \|w\| \left\|\mathrm{Vdm}(\tilde{\mathbf{m}}_1)^{-1}\right\| \left\|\tilde{V}\right\| \\
&\le \zeta^{-42k^2(1+\lg k)-20k-6}\pi_{\min}^{-5-2\lg k}\varepsilon.
\end{aligned}$$

■

Lemma 20 *The matrices* $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}$, *and* $\tilde{\mathbf{B}}'$ *satisfy*

$$\left\|\tilde{\mathbf{A}}^{-1}\right\|, \left\|\tilde{\mathbf{B}}^{-1}\right\|, \left\|(\tilde{\mathbf{B}}')^{-1}\right\| \leq \zeta^{-7k^2}.$$

Proof By Lemma 33 and Corollary 8. ■

Lemma 21 *The matrices* $\tilde{\mathbf{B}}'$ *and* $\tilde{\mathbf{B}}$ *produced by Algorithm 1 will satisfy*

$$\left\|\tilde{\mathbf{B}}' - \mathbf{B}'\right\|, \left\|\tilde{\mathbf{B}} - \mathbf{B}\right\| \leq \zeta^{-42k^2(1+\lg k)-14k^2-20k-16}\pi_{\min}^{-6-2\lg k}\varepsilon.$$

Proof We'll prove the claim for $\tilde{\mathbf{B}}$; the proof is identical for $\tilde{\mathbf{B}}'$. We can bound $\tilde{\mathbf{B}} - \mathbf{B}$ using the same tools as in the previous bounds. First, we bound $\left\|\tilde{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\| \left\|\tilde{\mathbf{A}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| \leq \zeta^{-7k^2-2}\pi_{\min}^{-1}\varepsilon$. Now

$$\left\|\tilde{\mathbf{C}}_{\mathcal{BA}}\right\| \left\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| \leq \zeta^{-4}\zeta^{-42k^2(1+\lg k)-14k^2-20k-11}\pi_{\min}^{-6-2\lg k}\varepsilon$$

and

$$\left\|\tilde{\mathbf{C}}_{\mathcal{BA}}\right\| \left\|\tilde{\mathbf{A}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1} - \pi_{\odot}^{-1}\right\| \leq \zeta^{-4}\zeta^{-7k^2}\zeta^{-42k^2(1+\lg k)-7k^2-16k-6}\pi_{\min}^{-5-2\lg k}\varepsilon.$$

The resulting bound is

$$\begin{aligned}
\left\|\tilde{\mathbf{B}} - \mathbf{B}\right\| &= \left\|\tilde{\mathbf{C}}_{\mathcal{BA}}(\tilde{\mathbf{A}}^{\mathsf{T}})^{-1}\pi_{\odot}^{-1} - \mathbf{C}_{\mathcal{BA}}(\mathbf{A}^{\mathsf{T}})^{-1}\pi_{\odot}\right\| \\
&\leq \left\|\tilde{\mathbf{C}}_{\mathcal{BA}} - \mathbf{C}_{\mathcal{BA}}\right\| \left\|\tilde{\mathbf{A}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| \\
&+ \left\|\tilde{\mathbf{C}}_{\mathcal{BA}}\right\| \left\|\tilde{\mathbf{A}}^{-1} - \mathbf{A}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| \\
&+ \left\|\tilde{\mathbf{C}}_{\mathcal{BA}}\right\| \left\|\tilde{\mathbf{A}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1} - \pi_{\odot}^{-1}\right\| \\
&\leq \zeta^{-42k^2(1+\lg k)-14k^2-20k-16}\pi_{\min}^{-6-2\lg k}\varepsilon.
\end{aligned}$$

■

Lemma 22 *Algorithm 1 will compute* $\tilde{\mathbf{m}}_i$ *satisfying*

$$\left\|\tilde{\mathbf{m}}_i - \mathbf{m}_i\right\|_{\infty} \leq \zeta^{-42k^2(1+\lg k)-14k^2-20k-19}\pi_{\min}^{-5-2\lg k}\varepsilon$$

for all i.

Proof We'll compute the bound using the inversion of $\tilde{\mathbf{B}}$ since this will give us the worst case. Let $\tilde{y} = (\tilde{\mathrm{g}}(B_1 \cup \{i\}), \ldots, \tilde{\mathrm{g}}(B_k \cup \{1\})$ and let $y = (\mathrm{g}(B_1 \cup \{i\}), \ldots, \mathrm{g}(B_k \cup \{1\})$. We note that $\|\tilde{y} - y\| \leq \zeta^{-1}\zeta^{-1}\varepsilon$ by assumption, and $\|\tilde{y}\| \leq \zeta^{-2}$. Then

$$\|\tilde{y} - y\| \left\|\tilde{\mathbf{B}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| \leq \zeta^{-7k^2-2}\pi_{\min}^{-2}\varepsilon,$$

$$\|\tilde{y}\| \left\|\tilde{\mathbf{B}}^{-1} - \mathbf{B}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| \leq \zeta^{-42k^2(1+\lg k)-14k^2-20k-19} \pi_{\min}^{-8-2\lg k} \varepsilon,$$

and

$$\|\tilde{y}\| \left\|\tilde{\mathbf{B}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1} - \pi_{\odot}^{-1}\right\| \leq \zeta^{-42k^2(1+\lg k)-16k-7k^2-3} \pi_{\min}^{-5-2\lg k} \varepsilon^2.$$

Of the three terms, the middle one clearly dominates so that we get

$$\begin{aligned}
\|\tilde{\mathbf{m}}_i - \mathbf{m}_i\| &= \left\|\tilde{y}\tilde{\mathbf{B}}^{-1}\tilde{\pi}_{\odot}^{-1} - y\mathbf{B}^{-1}\pi_{\odot}^{-1}\right\| \\
&\leq \|\tilde{y} - y\| \left\|\tilde{\mathbf{B}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| + \|\tilde{y}\| \left\|\tilde{\mathbf{B}}^{-1} - \mathbf{B}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1}\right\| + \|\tilde{y}\| \left\|\tilde{\mathbf{B}}^{-1}\right\| \left\|\tilde{\pi}_{\odot}^{-1} - \pi_{\odot}^{-1}\right\| \\
&\leq \zeta^{-42k^2(1+\lg k)-14k^2-20k-19} \pi_{\min}^{-8-2\lg k} \varepsilon
\end{aligned}$$

■

Appendix B. Proof of Theorem 5

We start with some definitions. Let $V_{(i)}$ be the subspace of $\mathbb{R}^k$ spanned by the i'th standard basis vector, and $P_{(i)}$ the projection onto $V_{(i)}$ w.r.t. usual inner product (i.e., as a matrix, all 0's except a 1 in entry (i,i)). For a separated vector $v = (v_1, \ldots, v_k)$ define the Lagrange interpolation polynomials $p_{v,i}$ by $p_{v,i}(v_j) = \delta_{ij}$. We then have the matrix equation

$$p_{v,i}(v_{\odot}) = P_{(i)}. \tag{14}$$

Write $p_{v,i}(x) = \sum_{j=0}^{k-1} p_{v,i,j} x^j$. Define the "coefficient norm" of a polynomial by $\|p_{v,i}\|_c = \sum_{j=0}^{k-1} j \cdot |p_{v,i,j}|$.

Definition 23 *For any subspace $U \subseteq \mathbb{R}^n$, let $\mathcal{P}_U$ denote the orthogonal projection onto U and let $\mathcal{P}_U^{\perp}$ denote the orthogonal complement subspace. Then $I = \mathcal{P}_U + \mathcal{P}_U^{\perp}$.*

In the following, $\|\cdot\|$ will always denote the $2 \to 2$ operator norm of a matrix.

Definition 24 (U-operator norm) *Given a subspace $U \subseteq \mathbb{R}^n$, we can define the U-operator norm of a matrix $C \in \mathbb{R}^{m \times n}$, denoted $\|C\|_U$, as follows:*

$$\|C\|_U \coloneqq \max_{0 \neq u \in U} \|Cu\| / \|u\|.$$

Lemma 25 *Let U be any subspace of $\mathbb{R}^n$ and let $C \in \mathbb{R}^{n \times n}$. For all $n \geq 1$, $\left\|\mathcal{P}_U^{\perp} C^n\right\|_U \leq n\|C\|^{n-1}\|\mathcal{P}_U^{\perp} C\|_U$.*

Proof We need to show that for all $n \geq 2$,

$$\max_{0 \neq u \in U} \left\|\mathcal{P}_U^{\perp} C^n u\right\| / \|u\| \leq n \|C\|^{n-1} \max_{0 \neq u \in U} \left\|\mathcal{P}_U^{\perp} C u\right\| / \|u\|.$$

Using that $\|\mathcal{P}_U\|, \left\|\mathcal{P}_U^{\perp}\right\| \leq 1$, we have

$$\left\|\mathcal{P}_U^{\perp} C^n u\right\| = \left\|\mathcal{P}_U^{\perp} C (\mathcal{P}_U + \mathcal{P}_U^{\perp}) C^{n-1} u\right\| \leq \left\|\mathcal{P}_U^{\perp} C \mathcal{P}_U C^{n-1} u\right\| + \left\|\mathcal{P}_U^{\perp} C \mathcal{P}_U^{\perp} C^{n-1} u\right\|.$$

So (using $\left\|C^{n-1}\right\| \le \|C\|^{n-1}$ and $\|\mathcal{P}_U\| \le 1$ in the first term, and $\left\|\mathcal{P}_U^\perp\right\| \le 1$ in the second term)

$$\max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp C^n u\right\| / \|u\| \le \|C\|^{n-1} \max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp C u\right\| / \|u\| + \|C\| \max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp C^{n-1} u\right\| / \|u\|$$

and applying induction this is

$$\le \left(\|C\|^{n-1} + (n-1)\|C\|^{n-1}\right) \max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp C u\right\| / \|u\| .$$

■

Recall from Section 3 that the coefficient norms of the interpolation polynomials are $\|p_{v,i}\|_c = \sum_0^{k-1} j\,|p_{v,i,j}|$.

Lemma 26 $\left\|\mathcal{P}_U^\perp P_{(i)}\right\|_U \le \|p_{v,i}\|_c \cdot \left\|\mathcal{P}_U^\perp v_\odot\right\|_U$.

Proof We are to show that $\max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp P_{(i)} u\right\| / \|u\| \le \|p_{v,i}\|_c \max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp v_\odot u\right\| / \|u\|$. We have

$$\max_u \left\|\mathcal{P}_U^\perp P_{(i)} u\right\| / \|u\| = \max_u \left\|\mathcal{P}_U^\perp p_{v,i}(v_\odot) u\right\| / \|u\| \le \sum_j \max_u |p_{v,i,j}| \cdot \left\|\mathcal{P}_U^\perp (v_\odot)^j u\right\| / \|u\| .$$

Note that $\|v_\odot\| \le 1$. So applying Lemma 25:

$$\max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp P_{(i)} u\right\| / \|u\| \le \sum_j |p_{v,i,j}|\, j \max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp v_\odot u\right\| / \|u\| = \|p_{v,i}\|_c \max_{0\neq u\in U} \left\|\mathcal{P}_U^\perp v_\odot u\right\| / \|u\| .$$

■

Lemma 27 *Let $\mathbb{1} \in U \subsetneq \mathbb{R}^k$ (strict containment). Then there is an i s.t. $\|\mathcal{P}_U^\perp P_{(i)} \mathbb{1}/\sqrt{k}\| \ge 1/k$.*

Proof Fixing any unit vector $v \in U^\perp$, we have for all i that $\|\mathcal{P}_U^\perp P_{(i)} \mathbb{1}/\sqrt{k}\| \ge \|v^* P_{(i)} \mathbb{1}/\sqrt{k}\|$. Select an i s.t. $|v_i| \ge \frac{1}{\sqrt{k}}$. Then $\|v^* P_{(i)} \mathbb{1}/\sqrt{k}\| \ge 1/k$. ■

Lemma 28 *Let $\mathbb{1} \in U \subset \mathbb{R}^k$. Then $\|\mathcal{P}_U^\perp v_\odot\|_U > \beta$.*

Proof By Lemma 27, there is an i s.t. $\left\|\mathcal{P}_U^\perp P_{(i)}\right\|_U > 1/k$. Applying Lemma 26 to this i, we have that $\left\|\mathcal{P}_U^\perp v_\odot\right\|_U \ge \frac{1}{k\|p_{v,i}\|_c}$. Now we need an upper bound on $\|p_{v,i}\|_c$. Recall $p_{v,i}(x) = \frac{\prod_{j\neq i}(x-\lambda_j)}{\prod_{j\neq i}(\lambda_i-\lambda_j)}$. Simply by upper-bounding all λ_j by 1 and lower-bounding all separations by ζ, we have the bound $\|p_{v,i}\|_c \le \zeta^{1-k} \sum_{\ell'=0}^{k-1} \ell' \binom{k-1}{\ell'} \le (k-1)(2/\zeta)^{k-1}$. ■

Lemma 29 *Fix ζ-separated vectors $\mathbf{m}_1, \ldots, \mathbf{m}_{k-1}$. Then there exist k row vectors $v_1, \ldots, v_k$, with $v_1 = \mathbb{1}/\sqrt{k}$, and for $\ell \ge 1$, $v_{\ell+1} \coloneqq \mathbf{m}_\ell \odot u$ where u is a unit vector in $U_\ell \coloneqq \operatorname{span}\{v_1, \ldots, v_\ell\}$, and such that*

- *Any unit vector in U_ℓ can be formed as a linear combination of the rows in $[\mathbf{m}_R]_{R\subseteq[\ell-1]}$ with coefficients bounded in maximum magnitude by $\beta^{1-\ell}$.*

Proof We prove the slightly tighter bound $\beta^{1-\ell}/\sqrt{k}$. We induct on ℓ. For $\ell = 1$ the bound is exact by construction. For $\ell > 1$, apply Lemma 28 to find a unit vector $u^* \in U_{\ell-1}$ such that $\left\|\mathcal{P}_U^{\perp}(\mathbf{m}_\ell \odot u^*)\right\| > \beta$. Set $v_\ell := \mathbf{m}_\ell \odot u^*$. Then $U_\ell = \operatorname{span}(U_{\ell-1}, \mathcal{P}_U^{\perp}(\mathbf{m}_\ell \odot u^*))$. The operator norm of $(\mathbf{m}_\ell)_\odot$ is ≤ 1 so $\|\mathbf{m}_\ell \odot u^*\| \leq 1$. Any unit vector $u \in U_\ell$ can be written as $c_1 u^{(\ell-1)} + c_2(\mathbf{m}_\ell \odot u^*)$ for $u^{(\ell-1)} \in U_{\ell-1}$ a unit vector and where $|c_2| \leq \beta^{-1}$ and $|c_1| \leq \beta^{-1}$. (This is just the operator norm of the inverse of $\begin{bmatrix} 1 & 0 \\ \sqrt{1-\beta^2} & \beta \end{bmatrix}$.) By the induction hypothesis, we can expand $u^{(\ell-1)}$ and u^* in terms of $\mathbf{m}_{R\subseteq[\ell-2]}$ as follows:

$$\begin{aligned} u &= c_1 u^{(\ell-1)} + c_2(\mathbf{m}_\ell \odot u^*) \\ &= c_1 \sum_{R\subseteq[\ell-2]} \alpha_R \mathbf{m}_R + c_2 \mathbf{m}_\ell \sum_{R\subseteq[\ell-2]} \alpha'_R \mathbf{m}_R \\ &= \sum_{R\subseteq[\ell-2]} \left(c_1 \alpha_R \mathbf{m}_R + c_2 \alpha'_R \mathbf{m}_{R\cup\{\ell\}}\right) \end{aligned}$$

where $\|\alpha\|_\infty, \|\alpha'\|_\infty \leq \beta^{2-\ell}/\sqrt{k}$. The claim follows immediately. ■

Corollary 30 *For any unit column vector $z \in \mathbb{R}^k$, the matrix $\mathbf{M}[2^S]$ satisfies $\left\|\mathbf{M}[2^S]z\right\|_\infty \geq \beta^k 2^{-k}$.*

Proof By Lemma 29, we know that we can write $z^\mathsf{T} = \lambda^\mathsf{T}\mathbf{M}$ where $\lambda \in \mathbb{R}^{2^k}$ and $\|\lambda\|_\infty \leq \beta^{-k}$. Thus, $1 = \|z\|^2 = \sum_{R\subseteq S} \lambda_R \mathbf{m}_R z$. There must be some $R \subseteq S$ for which $|\lambda_R \mathbf{m}_R z| \geq 1/2^k$. Since $|\lambda_R| \leq \beta^{-k}$ we immediately get that $\|\mathbf{M}z\|_\infty \geq |\mathbf{m}_R z| \geq \beta^k 2^{-k}$. ■

For a subset T with $|T| > k-1$, the bound on the largest singular value of $\mathbf{M}[2^T]$ increases to $k2^{|T|}$ while the lower bound remains unchanged.

Corollary 31 *$\sigma_{\max}(\mathbf{M}[2^S]) \leq k2^{k-1}$ and $\sigma_k(\mathbf{M}[2^S]) \geq \beta^k 2^{-k}/k$.*

Proof The largest singular value of $\mathbf{M}[2^S]$ is easily upper bounded by $k2^{k-1}$. Lemma 29 gives the bound on σ_k. ■

Bounding the condition number of a $k \times k$ submatrix of $\mathbf{M}[2^S]$. We can now use the following result from Feldman et al. (2008) to find a $k \times k$ submatrix that is similarly well-conditioned.

Lemma 32 (Corollary 6 in Feldman et al. (2008)) *Let $A \in \mathbb{R}^{k\times n}$ with $k < n$, and let $\sigma_k(A) \geq \varepsilon$. Then there exists a subset of the columns $J \subseteq [n]$ with $|J| = k$ such that $\sigma_k(A_J) \geq \varepsilon/\sqrt{k(n-k)+1}$.*

Proof of Theorem 5. The upper bound is trivial since all entries are in $[0, 1]$. The lower bound follows by applying Lemma 32 to Corollary 31.

Appendix C. Miscellaneous Proofs

Lemma 33 *For an invertible $n \times n$ matrix M and a perturbed matrix $\tilde{M}$, if $\left\|\tilde{M} - M\right\| = \varepsilon \leq \sigma_n(M)/2$, then*

$$\left\|\tilde{M}^{-1} - M^{-1}\right\| \leq 2\left\|M^{-1}\right\|^2 \varepsilon, \quad \text{and } \left\|\tilde{M}^{-1}\right\| \leq 2\left\|M^{-1}\right\|.$$

Proof First, we observe that

$$\left\|\tilde{M}^{-1}\right\| = \frac{1}{\sigma_n(\tilde{M})} \leq \frac{1}{\sigma_n(M) - \sigma_n(M)/2} \leq 2\left\|M^{-1}\right\|.$$

We use the identity $\tilde{M}^{-1} - M^{-1} = \tilde{M}^{-1}\left(M - \tilde{M}\right)M^{-1}$.

$$\left\|\tilde{M}^{-1} - M^{-1}\right\| = \left\|\tilde{M}^{-1}\left(M - \tilde{M}\right)M^{-1}\right\| \leq 2\left\|M^{-1}\right\|^2 \left\|M - \tilde{M}\right\|.$$

■