Recovering Structured Probability Matrices[*][†]

Qingqing Huang[1], Sham M. Kakade[2], Weihao Kong[3], and Gregory Valiant[4]

1 Massachusetts Institute of Technology, Cambridge, MA, USA
qqh@mit.edu
2 University of Washington, Seattle, WA, USA
am@cs.washington.edu
3 Stanford University, Stanford, CA, USA
whkong@stanford.edu
4 Stanford University, Stanford, CA, USA
valiant@stanford.edu

Abstract

We consider the problem of accurately recovering a matrix $\mathbb{B}$ of size $M \times M$, which represents a probability distribution over M^2 outcomes, given access to an observed matrix of "counts" generated by taking independent samples from the distribution $\mathbb{B}$. How can structural properties of the underlying matrix $\mathbb{B}$ be leveraged to yield computationally efficient and information theoretically optimal reconstruction algorithms? When can accurate reconstruction be accomplished in the sparse data regime? This basic problem lies at the core of a number of questions that are currently being considered by different communities, including building recommendation systems and collaborative filtering in the sparse data regime, community detection in sparse random graphs, learning structured models such as topic models or hidden Markov models, and the efforts from the natural language processing community to compute "word embeddings". Many aspects of this problem—both in terms of learning and property testing/estimation and on both the algorithmic and information theoretic sides—remain open.

Our results apply to the setting where $\mathbb{B}$ has a low rank structure. For this setting, we propose an efficient (and practically viable) algorithm that accurately recovers the underlying $M \times M$ matrix using $\Theta(M)$ samples (where we assume the rank is a constant). This linear sample complexity is optimal, up to constant factors, in an extremely strong sense: even testing basic properties of the underlying matrix (such as whether it has rank 1 or 2) requires $\Omega(M)$ samples. Additionally, we provide an even stronger lower bound showing that distinguishing whether a sequence of observations were drawn from the uniform distribution over M observations versus being generated by a well-conditioned Hidden Markov Model with two hidden states requires $\Omega(M)$ observations, while our positive results for recovering $\mathbb{B}$ immediately imply that $\Omega(M)$ observations suffice to *learn* such an HMM. This lower bound precludes sublinear-sample hypothesis tests for basic properties, such as identity or uniformity, as well as sublinear sample estimators for quantities such as the entropy rate of HMMs.

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases Random matrices, matrix recovery, stochastic block model, Hidden Markov Models

Digital Object Identifier 10.4230/LIPIcs.ITCS.2018.46

[*] Sham Kakade acknowledges funding from the Washington Research Foundation for innovation in Data-intensive Discovery. Gregory and Weihao's contributions were supported by NSF CAREER Award CCF-1351108, and a Sloan Research Fellowship.

[†] full version at https://arxiv.org/abs/1602.06586

9th Innovations in Theoretical Computer Science Conference (ITCS 2018).
Editor: Anna R. Karlin; Article No. 46; pp. 46:1–46:14
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Consider an unknown $M \times M$ matrix of probabilities $\mathbb{B}$, satisfying $\sum_{i,j} \mathbb{B}_{i,j} = 1$. Suppose one is given N independently drawn (i, j)-pairs, sampled according to the distribution defined by $\mathbb{B}$. How many draws are necessary to accurately recover $\mathbb{B}$? What can one infer about the underlying matrix based on these samples? How can one accurately test whether the underlying matrix possesses certain properties of interest? How do structural assumptions on $\mathbb{B}$ — for example, the assumption that $\mathbb{B}$ has low rank — affect the information theoretic or computational complexity of these questions? For the majority of these tasks, we currently lack both a basic understanding of the computational and information theoretic lay of the land, as well as algorithms that seem capable of achieving the information theoretic or computational limits.

This general question of making accurate inferences about a matrix of probabilities, given a matrix of observed "counts" of discrete outcomes, lies at the core of a number of problems that disparate communities have been tackling independently. On the theoretical side, these problems include both work on community detection in stochastic block models (where the goal is to infer the community memberships from an adjacency matrix of a graph that has been drawn according to an underlying matrix of probabilities expressing the community structure) as well as the line of work on recovering topic models, hidden Markov models (HMMs), and richer structured probabilistic models (where the model parameters can often be recovered using observed count data). On the practical side, these problems include work on computing low-rank approximations to sparsely sampled data, which arise in collaborative filtering and recommendation systems, as well as the recent work from the natural language processing community on understanding matrices of word co-occurrence counts for the purpose of constructing good "word embeddings". Additionally, work on latent semantic analysis and non-negative matrix factorization can also be recast in this setting.

In this work, we focus on this estimation problem where the probability matrix $\mathbb{B}$ possesses a particular low rank structure. While this estimation problem is rather specific, it generalizes the basic community detection problem and the problem of learning various common models encountered in natural language processing such as *probabilistic latent semantic analysis* [28]. Additionally, this problem encompasses the main technical challenge behind learning HMMs and topic models, in the sense that after $\mathbb{B}$ is accurately recovered, these learning problems have a number of parameters that is a function only of the number of topics/hidden states (which bounds the rank of $\mathbb{B}$ and is, in practical applications, at most a few hundred) as opposed to the the dictionary/alphabet size, M, which, in natural language settings is typically tens of thousands. Furthermore, this low rank case also provides a means to study how the relationships between property testing and estimation problems differ between this structured setting and the basic rank 1 setting that is equivalent to simply drawing i.i.d samples from a distribution supported on M elements.

We focus on the estimation of a low rank probability matrix $\mathbb{B}$ in the sparse data regime, near the information theoretic limit. In many practical scenarios involving sample counts, we seek algorithms capable of extracting the underlying structure in the sparsely sampled regime. To give two motivating examples, consider forming the matrix of word co-occurrences—the matrix whose rows and columns are indexed by the set of words, and whose (i, j)-th element consists of the number of times the i-th word follows the j-th word in a large corpus of text. In this context, the underlying probability matrix, $\mathbb{B}$, represents the distribution of bi-grams encountered in written english. In the context of recommendation system, one could consider a low rank matrix model, where the rows are indexed by customers, and the columns are indexed by products, with the (i, j)-th entry corresponding to the number of times the i-th customer has purchased the j-th product. Here, the underlying probability matrix, $\mathbb{B}$, models the distribution from which each customer/product purchase is drawn. In both settings, the structure of the probability matrix underlying these observed counts contains insights into the two domains, and in both domains we only have relatively sparse data. This is inherent in many other natural scenarios involving heavy-tailed distributions (including genomic settings), where despite having massive datasets, a significant fraction of the domain is observed only a single time.

Similar estimation questions have been actively studied in the community detection literature, where the objective is to accurately recover the communities in the regime where the average degree (e.g. the row sums of the adjacency matrix) are constant. In contrast, the recent line of works for recovering highly structured models (such as topic models, HMMs, etc.) are only applicable to the *over-sampled* regime where the amount of data is well beyond the information theoretic limits. In these cases, achieving the information theoretic limits remains a widely open question. This work begins to bridge the divide between these recent algorithmic advances in both communities. We hope that the low rank probability matrix setting considered here serves as a jumping-off point for the more general questions of developing information theoretically optimal algorithms for estimating structured matrices and tensors in general, or recovering low-rank approximations to arbitrary probability matrices, in the sparse data regime. While the general settings are more challenging, we believe that some of our algorithmic techniques can be fruitfully extended.

In addition to developing algorithmic tools which we hope are applicable to a wider class of problems, a second motivation for considering this particular low rank case is that, with respect to distribution learning and property testing, the entire lay-of-the-land seems to change completely when the probability matrix $\mathbb{B}$ has rank larger than 1. In the rank 1 setting — where a sample consists of 2 *independent* draws from a distribution supported on $\{1, \ldots, M\}$ — the distribution can be learned using $\Theta(M)$ draws. Nevertheless, many properties of interest can be tested or estimated using a sample size that is *sublinear* in M[1]. However, even just in the case where the probability matrix is of rank 2, although the underlying matrix $\mathbb{B}$ can be represented with $O(M)$ parameters (and, as we show, it can also be accurately and efficiently recovered with $O(M)$ sample counts), sublinear sample property testing and estimation is generally impossible. This result begs a more general question: *what conditions must be true of a structured statistical setting in order for property testing to be easier than learning?*

1.1 Problem Formulation

We consider the following problem setup and notation:

- A vocabulary consisting of M “words”, denoted by $\mathcal{M} = \{1, \ldots, M\}$.
- A low rank probability matrix $\mathbb{B}$, of size $M \times M$, with the following structure: $\mathbb{B} = \mathbb{P}\mathbb{W}\mathbb{P}^\top$, where $\mathbb{P}$ is an $M \times r$ non-negative matrix with column sums 1, and $\mathbb{W}$ is p.s.d. with $\sum_{i,j} \mathbb{W}_{i,j} = 1$.
- A set of N independent (i,j) pairs drawn according to $\mathbb{B}$, with the probability of drawing (i,j) given by $\mathbb{B}_{i,j}$.
- An $M \times M$ matrix of “counts”, C, summarizing the frequencies of each (i,j) pair in the N draws.

[1] Distinguishing whether a distribution is uniform versus far from uniform can be accomplished using only $O(\sqrt{M})$ draws, testing whether two sets of samples were drawn from similar distributions can be done with $O(M^{2/3})$ draws, estimating the entropy of the distribution to within an additive ϵ can be done with $O(\frac{M}{\epsilon \log M})$ draws, etc.

Throughout, we will make frequent use of the Poissonization technique whereby we assume that the number of draws follows a Poisson distribution of expectation N. This renders $C_{i,j}$ independent of the other entries of the count matrix, simplifying analysis. Additionally, for both upper and lower bounds, with all but inverse exponential probability the $o(N)$ discrepancy between N and $Poi(N)$ contributes only to lower order terms.

Notation

Throughout the paper, we use the following standard shorthand notations. Denote $[n] \triangleq \{1, \dots, n\}$. $\mathcal{I}$ denotes a subset of indices in $\mathcal{M}$. For a M-dimensional vector x, we use vector $x_{\mathcal{I}}$ to denote the elements of x restricted to the indices in $\mathcal{I}$; for two index sets $\mathcal{I}$, $\mathcal{J}$, and a $M \times M$ dimensional matrix X, we use $X_{\mathcal{I}\times\mathcal{J}}$ to denote the submatrix of X with rows restricting to indices in $\mathcal{I}$ and columns restricting to indices in $\mathcal{J}$.

We use $\mathrm{Poi}(\lambda)$ to denote a Poisson distribution with expectation λ; we use $\mathrm{Ber}(p)$ to denote a Bernoulli random variable with success probability $p \in [0,1]$; and for a probability vector $x \in [0,1]^M$ satisfying $\sum_i x_i = 1$ and an integer t, we use $\mathrm{Mul}(x;t)$ to denote the multinomial distribution over M outcomes corresponding to t draws from $[M]$ according to the distribution specified by the vector x.

1.2 Main Results

Our main result is the accurate recovery of a rank R matrix of the form described above in the linear data regime $N = O(M)$:

▶ **Theorem 1** (Upper bound for rank R, constant accuracy). *Suppose we have access to N i.i.d. samples generated according to the a probability matrix $\mathbb{B} = \mathbb{P}\mathbb{W}\mathbb{P}^T$ with $\mathbb{P}$ an $M \times R$ nonnegative matrix with column sum 1, $\mathbb{W}$ an $R \times R$ p.s.d. matrix with entries summing to 1 and row sums bounded by $\sum_j \mathbb{W}_{i,j} \geq w_{min}$. For any constants $\epsilon > 0, \delta > 0$ and $N = \Theta(\frac{MR^2}{w_{min}^2 \epsilon^5} \log(1/\delta))$, there is an algorithm with $poly(M, \log(1/\delta))$ runtime that returns a rank R matrix $\widehat{\mathbb{B}}$ such that with probability at least $1-\delta$:*

$$||\widehat{\mathbb{B}} - \mathbb{B}||_{\ell_1} \leq \epsilon.$$

We emphasize that our recovery is in terms of ℓ_1 distance, namely the total variation distance between the true distribution and the recovered distribution. In settings where there is a significant range in the row (or column) sums of $\mathbb{B}$, a spectral error bound might not be meaningful.

Much of the the difficulty in the algorithm is overcoming the fact that the row/column sums of $\mathbb{B}$ might be very non-uniform. Nevertheless, our result can be compared to the community detection setting with R communities (for which the row/column sums are completely uniform), for which accurate recovery can be efficiently achieved given $N = \Theta(MR^2)$ samples [20]. In our more general setting, we incur an extra factor of w_{min}^{-1}, whose removal might be possible with a more careful analysis of our approach.

1.2.1 Topic Models and Hidden Markov Models

One of the motivations for considering low rank structure of a probability matrix $\mathbb{B}$ is that this structure captures the structure of the matrix of expected bigrams generated by topic models [46, 28] and HMMs, as described below.

▶ **Definition 2.** An R-*topic model* over a vocabulary of size M is defined by a set of R distributions, $p^{(1)}, \dots, p^{(R)}$ supported over M words, and a set of R corresponding topic *mixing weights* $w_1, \dots, w_R$ with $\sum_i w_i = 1$. The process of drawing a bigram (i, j) consists of first randomly picking a topic $i \in [R]$ according to the distribution defined by the mixing weights, and then drawing two independent words from the distribution $p^{(i)}$ corresponding to the selected topic, i. Thus the probability of drawing a bigram (i, j) is $\sum_{k=1}^{R} w_R p^{(k)}(i) p^{(k)}(j)$, and the underlying distribution $\mathbb{B}$ over (i, j) pairs can be expressed as $\mathbb{B} = \mathbb{P}\mathbb{W}\mathbb{P}^\top$ with $\mathbb{P} = [p^{(1)}, \dots, p^{(R)}]$, and $\mathbb{W} = diag(w_1, \dots, w_R)$.

In the case of topic models, the decomposition of the matrix of bigram probabilities $\mathbb{B} = \mathbb{P}\mathbb{W}\mathbb{P}^\top$ has the desired form required by our Theorem 1, with $\mathbb{W}$ nonnegative and p.s.d., and hence the theorem guarantees an accurate recovery of $\mathbb{B}$, even in the sparse data regime. The recovery of the mixing weights $\{w_i\}$ and topic distributions $\{p^{(i)}\}$ from $\mathbb{B}$ requires an additional step, which will amount to solving a system of quadratic equations. Crucially, however, given the rank R matrix $\mathbb{B}$, the remaining problem becomes a problem only involving R^2 parameters—representing a linear combination of the R factors of $\mathbb{B}$ for each $p^{(i)}$—rather than recovering MR parameters.

▶ **Definition 3.** A *Hidden Markov model* with R hidden states and observations over an alphabet of size M is defined by an $R \times R$ transition matrix T, and R observation distributions $p^{(1)}, \dots, p^{(R)}$. A sequence of observations is sampled as follows: select an initial state (e.g. according to the stationary distribution of the chain) then evolve the Markov chain according to the transition matrix T, drawing an observation from the ith distribution $p^{(i)}$ at each timestep in which the underlying chain is in state ith.

Assuming the Markov chain has stationary distribution $\pi_1, \dots, \pi_R$, the probability of seeing a bigram (i, j) with symbol i observed at the kth timestep and symbol j observed at the $k+1$st timestep, tends towards the following (i.e. assuming the chain is close to mixing by timestep k) rank R probability matrix $\mathbb{B} = \mathbb{P}\mathbb{W}\mathbb{P}^\top$, with $\mathbb{P} = [p^{(1)}, \dots, p^{(R)}]$ and $\mathbb{W} = diag(\pi_1, \dots, \pi_n)T$.

For HMMs, the low rank matrix of bigrams, $\mathbb{B} = \mathbb{P}\mathbb{W}\mathbb{P}^\top$, does *not* necessarily have the required form—specifically the mixing matrix $\mathbb{W}$ may not be p.s.d.—and it is unclear whether our approach can successfully recover such matrices. Nevertheless, with slightly more careful analysis, at least in certain cases the techniques yield tight results. For example, in the setting of an HMM with two hidden states, over an alphabet of size M, we can easily show that our techniques obtain an accurate reconstruction of the corresponding probability matrix $\mathbb{B}$, and then leverage that reconstruction together with a constant amount of tri-gram information to accurately learn the HMM:

▶ **Proposition 4.** *(Learning 2-state HMMs) Consider a sequence of observations given by a Hidden Markov Model with two hidden states and symmetric transition matrix with entries bounded away from 0. Assuming a constant ℓ_1 distance between the distributions of observations corresponding to the two states, there exists an algorithm which, given a sampled chain of length $N = \Omega(M/\epsilon^2)$, runs in time poly(M) and returns estimates of the transition matrix and two observation distributions that are accurate in ℓ_1 distance, with probability at least* $2/3$.

This probability of failure can be trivially boosted to $1 - \delta$ at the expense of an extra factor of $\log(1/\delta)$ observations.

1.2.2 Testing vs. Learning

Theorem 1 and Proposition 4 are tight in an extremely strong sense: for both the topic model and HMM settings, it is information theoretically impossible to perform even the most basic property tests using fewer than $\Theta(M)$ samples. For topic models, the community detection lower bounds [43][34][55] imply that $\Theta(M)$ bigrams are necessary to even distinguish between the case that the underlying model is the uniform distribution over bigrams versus the case of a R-topic model in which each topic has a unique subsets of M/R words with a constant fraction higher probability than the remaining words. More surprisingly, for k-state HMMs with $k \geq 2$, even if we permit an estimator to have more information than merely bigram counts, namely access to the *full sequence* of observations, we prove the following linear lower bound.

▶ **Theorem 5.** *There exists a constant $c > 0$ such that for sufficiently large M, given a sequence of observations from a HMM with two states and emission distributions p, q supported on M elements, even if the underlying Markov process is symmetric, with transition probability $1/4$, it is information theoretically impossible to distinguish the case that the two emission distributions, $p = q =$ Unif$[M]$ from the case that $||p - q||_1 = 1$ with probability greater than $2/3$ using a sequence of fewer than cM observations.*

This immediately implies the following corollary for estimating the *entropy rate* of an HMM.

▶ **Corollary 6.** *There exists an absolute constant $c > 0$ such that given a sequence of observations from a HMM with two hidden states and emission distributions supported on M elements, a sequence of cM observations is information theoretically necessary to estimate the entropy rate to within an additive 0.5 with probability of success greater than $2/3$.*

These strong lower bounds for property testing and estimation are striking for several reasons. First, the core of our learning algorithm for 2-state HMMs (Proposition 4) is a matrix reconstruction step that uses only the set of bigram counts. Conceivably, it might be helpful to consider longer sequences of observations — even for HMMs that mix in constant time, there are detectable correlations between observations separated by $O(\log M)$ steps. Regardless, our lower bound shows that actually no additional information from such longer k-grams can be leveraged to yield sublinear sample property testing or estimation.

A second notable point is the apparent brittleness of sublinear property testing and estimation as we deviate from the standard (unstructured) i.i.d sampling setting. Indeed for nearly all distributional property estimation or testing tasks, including testing uniformity and estimating the entropy, sublinear-sample testing and estimation is possible in the i.i.d. sampling setting (e.g. [26, 52, 51]). In contrast to the i.i.d. setting in which estimation and testing require asymptotically fewer samples than *learning*, as the above results illustrate, even in the setting of an HMM with just two hidden states, learning and testing require comparable numbers of observations.

1.3 Related Work

As mentioned earlier, the general problem of reconstructing an underlying matrix of probabilities given access to a count matrix drawn according to the corresponding distribution, lies at the core of questions that are being actively pursued by several different communities. We briefly describe these questions, and their relation to the present work.

Community Detection. With the increasing prevalence of large scale social networks, there has been a flurry of activity from the algorithms and probability communities to both model structured random graphs, and understand how (and when it is possible) to examine a graph and infer the underlying structures that might have given rise to the observed graph. One of the most well studied community models is the *stochastic block model* [29]. In its most basic form, this model is parameterized by a number of individuals, M, and two probabilities, α, β. The model posits that the M individuals are divided into two equal-sized "communities", and such a partition defines the following random graph model: for each pair of individuals in the same community, the edge between them is present with probability α (independently of all other edges); for a pair of individuals in different communities, the edge between them is present with probability $\beta < \alpha$. Phrased in the notation of our setting, the adjacency matrix of the graph is generated by including each potential edge (i, j) independently, with probability $\mathbb{B}_{i,j}$, with $\mathbb{B}_{i,j} = \alpha$ or β according to whether i and j are in the same community. Note that $\mathbb{B}$ has rank 2 and is expressible as $\mathbb{B} = PWP^\top$ where $P = [p, q]$ for vectors $p = \frac{2}{M} I_1$ and $q = \frac{2}{M} I_2$ where I_1 is the indicator vector for membership in the first community, and I_2 is defined analogously, and W is the 2×2 matrix with $\alpha \frac{M^2}{4}$ on the diagonal and $\beta \frac{M^2}{4}$ on the off-diagonal.

What values of α, β, and M enable the community affiliations of all individuals to be accurately recovered with high probability? What values of α, β, and M allow for the graph to be distinguished from an Erdos-Renyi random graph (that has no community structure)? The crucial regime is where $\alpha, \beta = O(\frac{1}{M})$, and hence each person has a constant, or logarithmic expected degree. The naive spectral approaches will fail in this regime, as there will likely be at least one node with degree $\approx \log M / \log \log M$, which will ruin the top eigenvector. Nevertheless, in a sequence of works sparked by the paper of Friedman, and Szemeredi [24], the following punchline has emerged: the naive spectral approach will work, even in the constant expected degree setting, provided one first either removes, or at least diminishes the weight of these high-degree problem vertices (e.g. [23, 33, 42, 34, 35]). For both the *exact* recovery problem and the detection problem, the exact tradeoffs between α, β, and M were recently established, down to subconstant factors [43, 1, 38]. More recently, there has been further research investigating more complex stochastic block models, consisting of three or more components, components of unequal sizes, etc. (see e.g. [20, 2, 3]).

The community detection setting generates an adjacency matrix with entries in $\{0, 1\}$, choosing entry $C_{i,j} \leftarrow \mathit{Bernoulli}(\mathbb{B}_{i,j})$, as opposed to our setting where $C_{i,j}$ is drawn from the corresponding Poisson distribution. Nevertheless, the two models are extremely similar in the sparse regime considered in the community detection literature, since, when $\mathbb{B}_{i,j} = O(1/M)$, the corresponding Poisson and Bernoulli distributions have total variation distance $O(1/M^2)$.

Word Embeddings. On the more applied side, some of the most impactful advances in natural language processing over the past five years has been work on "word embeddings" [39, 37, 49, 10]. The main idea is to map every word w to a vector $v_w \in \mathbb{R}^d$ (typically $d \approx 500$) in such a way that the geometry of the vectors captures the semantics of the word.[2] One of the main constructions for such embeddings is to form the $M \times M$ matrix whose rows/columns are indexed by words, with (i, j)-th entry corresponding to the total number of times the i-th and j-th word occur next to (or near) each other in a large corpus of text (e.g. wikipedia).

[2] The goal of word embeddings is not just to cluster similar words, but to have semantic notions encoded in the geometry of the points: the example usually given is that the direction representing the difference between the vectors corresponding to "king" and "queen" should be similar to the difference between the vectors corresponding to "man" and "woman", or "uncle" and "aunt", etc.

The word embedding is then computed as the rows of the singular vectors corresponding to the top rank d approximation to this empirical count matrix.[3] These embeddings have proved to be extremely effective, particularly when used as a way to map text to features that can then be trained in downstream applications. Despite their successes, current embeddings seem to suffer from sampling noise in the count matrix (where many transformations of the count data are employed, e.g. see [48])—this is especially noticeable in the relatively poor quality of the embeddings for relatively rare words. The theoretical work [11] sheds some light on why current approaches are so successful, yet the following question largely remains: Is there a more accurate way to recover the best rank-d approximation of the underlying matrix than simply computing the best rank-d approximation for the (noisy) matrix of empirical counts?

Efficient Algorithms for Latent Variable Models. There is a growing body of work from the algorithmic side (as opposed to information theoretic) on how to recover the structure underlying various structured statistical settings. This body of work includes work on learning HMMs [31, 41, 19], recovering low-rank structure [9, 8, 15], and learning or clustering various structured distributions such as Gaussian mixture models [21, 54, 40, 14, 30, 32, 25]. A number of these methods essentially can be phrased as solving an inverse moments problem, and the work in [7] provides a unifying viewpoint for computationally efficient estimation for many of these models under a tensor decomposition perspective. In general, this body of work has focused on the computational issues and has considered these questions in the regime in which the amount of data is plentiful—well above the information theoretic limits.

On the practical side, the natural language processing community has considered a variety of generative and probabilistic models that fall into the framework we consider. These include work on *probabilistic latent semantic analysis* (see e.g. [28, 22]), including the popular *latent Dirichlet allocation* topic model [18]. Much of the algorithmic work on recovering these models is either of a heuristic nature (such as the EM framework), or focuses on computational efficiency in the regime in which data is plentiful (e.g. [6].

Sublinear Sample Testing and Estimation. In contrast to the work described in the previous section on efforts to devise computationally efficient algorithms for tackling complex structural settings in the "over–sampled" regime, there is also significant work establishing information theoretically optimal algorithms and (matching) lower bounds for estimation and distributional hypothesis testing in the most basic setting of independent samples drawn from (unstructured) distributions. This work includes algorithms for estimating basic statistical properties such as entropy [45, 27, 50, 52], support size [47, 50], distance between distributions [50, 52, 51], and various hypothesis tests, such as whether two distributions are very similar, versus significantly different [26, 12, 44, 53, 16], etc. While many of these results are optimal in a worst-case ("minimax") sense, there has also been recent progress on instance optimal (or "competitive") estimation and testing, e.g. [4, 5, 53], with stronger information theoretic optimality guarantees. There has also been a long line of work beginning with [17, 13] on these tasks in "simply structured" settings, e.g. where the domain of the distribution has a total ordering or where the distribution is monotonic or unimodal.

[3] A number of pre-processing steps have been considered, including taking the element-wise square roots of the entries, or logarithms of the entries, prior to computing the SVD.

2 Recovery Algorithm

To motivate our algorithms, it will be helpful to first consider the more naive approaches. Recall that we are given N samples drawn according to the probability matrix $\mathbb{B}$, with C denoting the matrix of empirical counts. By the Poisson assumption on sample size, we have that $C_{i,j} \sim \text{Poi}(N\mathbb{B}_{i,j})$. Perhaps the most naive hope is to consider the rank R truncated SVD of the empirical matrix $\frac{1}{N}C$, which concentrates to $\mathbb{B}$ in Frobenius norm at $\frac{1}{\sqrt{N}}$ rate. Unfortunately, in order to achieve constant ℓ_1 error, this approach would require a sample complexity as large as $\Theta(M^2)$. Intuitively, this is because the rows and columns of C corresponding to words with larger marginal probabilities have higher row and column sums in expectation, as well as higher variances that undermine the spectral concentration of the matrix as a whole.

The above observation leads to the idea of pre-scaling the matrix so that every word (i.e. row/column) roughly has equal variance. Indeed, with the pre-scaling modification of the truncated SVD, one can likely improve the sample complexity of this approach to $\Theta(M \log M)$. To further reduce the sample complexity, it is worth considering what prevents the truncated SVD from achieving accurate recovery in the $N = \Theta(M)$ regime. Suppose the word marginals are roughly uniform, namely all in the order of $O(\frac{1}{M})$, the linear sample regime roughly corresponds to the stochastic block model setup where the expected row sums are all of order $d = \frac{N}{M} = \Omega(1)$. It is well-known that in this sparse regime, the adjacency matrix (in the graph setting), or the empirical count matrix C in our problem, does not concentrate to the expectation matrix in the spectral sense. Due to heavy rows/columns of sum $\Omega(\frac{\log M}{\log \log M})$, the leading eigenvectors are polluted by the local properties of these heavy rows/columns and do not reveal the global structure of the matrix/graph, which is precisely the desired information.

Fortunately, these heavy (empirical) rows/columns are the *only* impediment to spectral concentration in the linear sample size regime. Provided all rows/columns with observed weight significantly more than d are zeroed out, spectral concentration prevails. This simple idea of taming the heavy rows/columns was first introduced by [24], and analyzed in [23] and many other works. Recently in [35] and [36], the authors provided clean and clever proofs to show that *any* manner of "regularization"—removing entries from the heavy rows/columns until their row/column sums are bounded—essentially leads to the desired spectral concentration for the adjacency matrix of random graphs whose row/column sums are roughly uniform in expectation.

The challenge of applying this regularization approach in our more general setting is that the row/column expectations of C might be extremely non-uniform. If we try to "regularize", we will not know whether we are removing entries from rows that have small expected sum but happened to have a few extra entries, or if we are removing entries from a row that actually has a large expected sum (in which case such removal will be detrimental).

Our approach is to partition the vocabulary $\mathcal{M}$ into bins that have roughly uniform marginal probabilities, corresponding to partitioning the rows/columns into sets that have roughly equal (empirical) counts. Restricting our attention to the diagonal sub-blocks of $\mathbb{B}$ whose rows/columns consist of indices restricted to a single bin, the expected row and column sums are now roughly uniform. We can regularize (by removing abnormally heavy rows and columns) from each diagonal block separately to restore spectral concentration on each of these sub blocks. Now, we can apply truncated SVD to each diagonal sub block, recovering the column span of these blocks of $\mathbb{B}$. With the column spans of each bin, we can now "stitch" them together as a single large projection matrix P which has rank at most R

Algorithm 1: The algorithm to which Theorem 1 applies, which recovers rank R probability matrices in the linear data regime.

Input: $3N$ i.i.d. samples from the distribution $\mathbb{B}$ of dimension $M \times M$, where $N = O(\frac{MR^2}{w_{min}^2 \epsilon^5})$

(In each of the 3 steps, B refers to an independent copy of the normalized count matrix $\frac{1}{N}C$.)

Output: Rank R estimator $\widehat{\mathbb{B}}$ for $\mathbb{B}$

Step 1. (**Binning according to the empirical marginal probabilities**)

Set $\widehat{\rho}_i = \frac{\sum_{j=1}^{M}(C_{i,j}+C_{j,i})}{2N}$. Partition the vocabulary $\mathcal{M}$ into:

$$\mathcal{I}_0 = \left\{ i : \widehat{\rho}_i < \frac{1}{N} \right\}, \text{ and } \mathcal{I}_k = \left\{ i : \frac{e^{k-1}}{N} \leq \widehat{\rho}_i \leq \frac{e^k}{N} \right\}, \text{ for } k = 1, \dots, \log N.$$

Sort the M words according to $\widehat{\rho}_i$ in ascending order. Define $\bar{\rho}_k = \frac{e^{k+1}}{N}$. For each bin $\mathcal{I}_k$, if $|\mathcal{I}_k| < 20e^{-\frac{3}{2}(k+1)}N$ set $\bar{\rho}_k$ to be 0. Let $k_0 = 4\log(\frac{c_0 R}{\epsilon\sqrt{w_{min}}}) + 16$, for an absolute constant c_0 which will be specified in the analysis, and set $\bar{\rho}_k$ to be 0 for all $k < k_0$. Define the following block diagonal matrix:

$$D = \begin{bmatrix} \bar{\rho}_1^{-1/2} I_{|\mathcal{I}_1|} & & \\ & \ddots & \\ & & \bar{\rho}_{\log N}^{-1/2} I_{|\mathcal{I}_{\log N}|} \end{bmatrix}. \qquad (1)$$

Step 2. (**Estimate dictionary span in each bin**)

For each diagonal block $B_k = B_{\mathcal{I}_k \times \mathcal{I}_k}$, perform the following two steps:

1. (**Regularization**):
   - If a row/column of B has sum exceeding $2\bar{\rho}_k$, set the entire row/column to 0.
   - If a row/column of B_k has sum exceeding $\frac{2|\mathcal{I}_k|\bar{\rho}_k^2}{w_{min}}$, set the entire row/column to 0.

   Denote the regularized block by $\widetilde{B}_k$.
2. (R-**SVD**): Define the $|\mathcal{I}_k| \times R$ matrix V_k to consist of the R top singular vectors of $\widetilde{B}_k$.

Step 3. (**Recover estimate for $\widehat{\mathbb{B}}$ accurate in ℓ_1**)

Define the following projection matrix:

$$P_V = \begin{bmatrix} P_{V_1} & & \\ & \ddots & \\ & & P_{V_{\log M}} \end{bmatrix}, \text{ where } P_{V_k} = V_k V_k^T. \qquad (2)$$

Let $\widehat{\mathbb{B}}'$ be the rank-R truncated SVD of matrix $P_V D^{-1} B D^{-1} P_V$, and return $\widehat{\mathbb{B}} = D\widehat{\mathbb{B}}'D$.

times the number of bins, and roughly contains the column span of $\mathbb{B}$. We then project a new count matrix, C', obtained via a fresh partition of samples. As the projection is fairly low rank, it filters most of the sampling noise, leaving an accurate approximation of $\mathbb{B}$.

We summarize these basic ideas of Algorithm 1.

1. Given a batch of N samples, group words according to the empirical marginal probabilities,

so that in each bin consists of words whose (empirical) marginal probabilities, differ by at most a constant factor.

2. Given a second batch of N samples, zeros out the words that have abnormally large empirical marginal probabilities comparing to the expected marginal probabilities of words in their bin. Then consider the diagonal blocks of the empirical bigram counts matrix C, with rows and columns corresponding to the words in the same bin. We "regularize" each diagonal block in the empirical matrix by removing abnormally heavy rows and columns of the blocks, and then apply truncated SVD to estimate the column span of that diagonal block of $\mathbb{B}$.
3. With a third batch of N samples, project the empirical count matrix into the "stitched" column spans recovered in the previous step which yields an accurate estimate of $\mathrm{Diag}(\rho)^{-1/2}\mathbb{B}\mathrm{Diag}(\rho)^{-1/2}$ in spectral norm, where ρ denotes the vector of marginal probabilities. Since the estimate is accurate in spectral norm *after* scaling by the marginal probabilities, this spectral concentration of the scaled matrix easily translates into an ℓ_1 error bounds for the un-scaled matrix $\mathbb{B}$, as desired.

There are several potential concerns that arise in implementing the above high-level algorithm outline and establishing the correctness of the algorithm:

1. We do not have access to the exact marginal probabilities of each word. With a linear sample size, the recovered vector of marginal probabilities has only constant (expected) accuracy in ℓ_1 norm. Hence each bin, defined in terms of the empirical marginals, includes some non-negligible fraction of words with significantly larger (or smaller) marginal probabilities. When directly applied to the empirical bins with such "spillover" words, the existing results of "regularization" in [36] do not lead to the desired concentration result.
2. When we restrict our analysis to a diagonal block corresponding to a single bin, we throw away all the sample counts outside of that block. This greatly reduces the effective sample size, since a significant fraction of a word's marginal probability might be due to co-occurrences with words outside of its bin. It is not obvious that we retain enough samples in each diagonal block to guarantee meaningful estimation. [If the mixing matrix $\mathbb{W}$ in $\mathbb{B} = \mathbb{P}\mathbb{W}\mathbb{P}^\top$ is not p.s.d., this effect may be sufficiently severe so as to render these diagonal blocks essentially empty, foiling this approach.]
3. Finally, even if the "regularization" trick works for each diagonal block, we need to extract the useful information and "stitch" together this information from each block to provide an estimator for the entire matrix, including the off-diagonal blocks. Fortunately, the p.s.d assumption of the mixing matrix W ensures that sufficient information is contained in these diagonal blocks.

References

1 Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

2 Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *CoRR*, abs/1503.00609, 2015. arXiv:1503.00609.

3 Emmanuel Abbe and Colin Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *CoRR*, abs/1512.09080, 2015. arXiv:1512.09080.

4 J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. In *Conference on Learning Theory (COLT)*, 2011.

5 J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive classification and closeness testing. In *Conference on Learning Theory (COLT)*, 2012.

6 Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 926–934, 2012. URL: `http://papers.nips.cc/paper/4637-a-spectral-algorithm-for-latent-dirichlet-allocation`.

7 Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014. URL: `http://jmlr.org/papers/v15/anandkumar14b.html`.

8 Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 145–162. ACM, 2012.

9 Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.

10 Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520, 2015. `arXiv:1502.03520`.

11 Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. *CoRR*, abs/1502.03520, 2015. `arXiv:1502.03520`.

12 T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1), 2013.

13 T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Symposium on Theory of Computing (STOC)*, pages 381–390, 2004.

14 Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

15 Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 594–603. ACM, 2014.

16 B. Bhattacharya and G. Valiant. Testing closeness with unequal sized samples. In *Neural Information Processing Systems (NIPS)*, 2015.

17 L. Birge. Estimating a density under order restrictions: Nonasymptotic minimax risk. *Annals of Statistics*, 15(3):995–1012, 1987.

18 David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

19 J. T. Chang. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.

20 Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery. *CoRR*, abs/1501.05021, 2015. `arXiv:1501.05021`.

21 Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

22 Chris H. Q. Ding, Tao Li, and Wei Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method.

In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 342–347. AAAI Press, 2006. URL: http://www.aaai.org/Library/AAAI/2006/aaai06-055.php.

23 Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

24 Joel Friedman, Jeff Kahn, and Endre Szemeredi. On the second eigenvalue of random regular graphs. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, pages 587–598. ACM, 1989.

25 Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Symposium on Theory of Computing, STOC 2015,*, 2015.

26 O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. In *Technical Report TR00-020, Electronic Colloquium on Computational Complexity*, 2000.

27 S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

28 Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

29 Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

30 Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

31 Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

32 Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

33 Raghunandan H Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, pages 324–328. IEEE, 2009.

34 Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

35 Can Le, Elizaveta Levina, and Roman Vershynin. Sparse random graphs: regularization and concentration of the laplacian. *arXiv preprint arXiv:1502.03049*, 2015.

36 Can M. Le and Roman Vershynin. Concentration and regularization of random graphs. *CoRR*, abs/1506.00669, 2015. arXiv:1506.00669.

37 Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2177–2185, 2014. URL: http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization.

38 Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

39 Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. arXiv:1301.3781.

40 Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

41 E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. *Annals of Applied Probability*, 16(2):583–614, 2006.

42 Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012. `arXiv:1202.1499`.

43 Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *CoRR*, abs/1407.1591, 2014. `arXiv:1407.1591`.

44 S. on Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1193–1203, 2014.

45 L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004.

46 Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM, 1998.

47 S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

48 Karl Stratos, Michael Collins, and Daniel Hsu. Model-based word embeddings from decompositions of count matrices. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, 2015.

49 Karl Stratos, Michael Collins Do-Kyum Kim, and Daniel Hsu. A spectral algorithm for learning class-based n-gram models of natural language. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.

50 G. Valiant and P. Valiant. Estimating the unseen: an $n/\log n$-sample estimator for entropy and support size, shown optimal via new clts. In *Symposium on Theory of Computing (STOC)*, 2011.

51 G. Valiant and P. Valiant. The power of linear estimators. In *Symposium on Foundations of Computer Science (FOCS)*, 2011.

52 G. Valiant and P. Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Neural Information Processing Systems (NIPS)*, 2013.

53 G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2014.

54 Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

55 Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *CoRR*, abs/1507.05313, 2015. URL: `http://arxiv.org/abs/1507.05313`, `arXiv:1507.05313`.