Recovering Structured Probability Matrices

Qingqing Huang* Sham M. Kakade† Weihao Kong‡ Gregory Valiant§

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 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 rank 2 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. This result easily translates to $\Theta(M)$ sample algorithms for learning topic models with two topics over dictionaries of size M, and learning hidden Markov Models with two hidden states and observation distributions supported on M elements. These linear sample complexities are 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. Furthermore, we provide an even stronger lower bound where distinguishing whether a sequence of observations were drawn from the uniform distribution over M observations versus being generated by an HMM with two hidden states requires $\Omega(M)$ observations. This 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. This impossibility of sublinear-sample property testing in these settings is intriguing and underscores the significant differences between these structured settings and the standard setting of drawing i.i.d samples from an unstructured distribution of support size M.

*MIT. Email:qqh@mit.edu.

†University of Washington. Email: sham@cs.washington.edu

‡Stanford University. Email:kweihao@gmail.com

§Stanford University. Email:valiant@stanford.edu

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 drawn according to an underlying matrix 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 the recent work in the natural language processing community to infer structure from matrices of word co-occurrence counts for the purpose of constructing good "word embeddings", as well as latent semantic analysis and non-negative matrix factorization.

In this work, we start this line of inquiry by focusing on the estimation problem where the probability matrix $\mathbb{B}$ possesses a rank 2 structure. While this estimation problem is rather specific, it generalizes the basic community detection problem and also encompasses the underlying problem behind learning 2-state HMMs and learning 2-topic models. Furthermore, this rank 2 case also provides a means to study how inferential and testing problems are different in this structured setting as opposed to the simpler rank 1 setting; the latter is equivalent to the standard setting of 2 independent draws from a distribution supported on M elements.

We focus on the estimation of $\mathbb{B}$ at the information theoretic limit. The motivation for this is that in many practical scenarios involving count samples, we seek algorithms capable of extracting the underlying structure in the sparsely sampled regime. To give two 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 ith word follows the jth word in a large corpus of text. One could also consider a matrix of genetic mutation co-occurrences, whose rows and columns are indexed by known mutations and whose (i, j)th entry is the number of individuals that have both mutations. 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, where regardless of how much data one collects, a significant fraction of items (e.g. words, genetic mutations) will only be observed a few times.

Such 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) is a constant. In contrast, the recent line of works for recovering highly structured models (such as large topic models, HMMs, etc.) are only applicable to the *over-sampled* regime (the amount of data is well beyond the information theoretic limits), where 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. In particular, the rank-2 probability matrix $\mathbb{B}$ can be efficiently recovered with a number of sample counts that is *linear* in M. The next questions are how to develop information theoretically optimal algorithms for estimating low rank

matrices and tensors in general, which may be a far more challenging setting; we hope that some of our algorithmic techniques can be extended here.

In addition to developing algorithmic tools which we hope are applicable to a wider class of problems, a second motivation for considering this rank 2 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, in the rank 2 setting, even though the underlying matrix $\mathbb{B}$ can be represented with $O(M)$ parameters (and, as we show, it can also be accurately recovered with $O(M)$ sample counts), sublinear sample property testing and estimation is generally impossible. This result begs the 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

Assume our vocabulary is the index set $\mathcal{M} = \{1, \ldots, M\}$ of M words and that there is an underlying low rank probability matrix $\mathbb{B}$, of size $M \times M$, with the following structure:

$$\mathbb{B} = DWD^\top, \text{ where } D = \Big[p, q\Big]. \tag{1}$$

Here, $D = \mathbb{R}_+^{M \times 2}$ is the *dictionary matrix* parameterized by two M-dimensional probability vectors p, q, supported on the standard $(M-1)$-simplex. Also, W is the 2×2 *mixing matrix*, which is a probability matrix satisfying $W \in \mathbb{R}_+^{2\times 2}, \sum_{i,j} W_{i,j} = 1$.

Define the *covariance matrix* of any probability matrix P as:

$$[\mathrm{Cov}(P)]_{i,j} := P_{i,j} - (\sum_k P_{i,k})(\sum_k P_{k,j}).$$

Note that $\mathrm{Cov}(P)\vec{1} = \vec{0}$ and $\vec{1}^\top \mathrm{Cov}(P) = \vec{0}$ (where $\vec{1}$ and $\vec{0}$ are the all ones and zeros vectors, respectively). This implies that, without loss of generality, the covariance of the mixing matrix, $\mathrm{Cov}(W)$, can be expressed as:

$$\mathrm{Cov}(W) = [w_L, -w_L]^\top [w_R, -w_R].$$

for some real numbers $w_L, w_R \in [-1, 1]$. For ease of exposition, we restrict to the symmetric case where $w_L = w_R = w$, though our results hold more generally.

Suppose we obtain N, i.i.d. sample counts from $\mathbb{B}$ of the form $\{(i_1, j_1)\,(i_2, j_2), \ldots (i_N, j_N)\}$, where each sample $(i_n, j_n) \in \mathcal{M} \times \mathcal{M}$. The probability of obtaining a count (i, j) in a sample is $\mathbb{B}_{i,j}$. Moreover, assume that the number of samples follows a Poisson distribution: $N \sim \mathrm{Poi}(\mathbb{N})$. The Poisson assumption on the number of samples is made only for the convenience of analysis: if N follows a Poisson distribution, the counts of observing (i, j) follows a Poisson distribution $\mathrm{Poi}(\mathbb{N}\mathbb{B}_{i,j})$ and is independent from the counts of observing (i', j') for $(i', j') \neq (i, j)$. This assumption is made only for the convenience of analysis and is not crucial for the correctness of the algorithm. As M is asymptotically large, with high probability, N and $\mathbb{N}$ are within a subconstant factor of each other and both upper and lower bounds translate between the Poissonized setting, and the setting of exactly N samples. Throughout, we state our sample complexity results in terms of N rather than $\mathbb{N}$.

[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.

Notation Throughout the paper, we use the following standard shorthand notations.

Denote $[n] \triangleq \{1, \ldots, n\}$. Let $\mathcal{I}$ denote 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 restricting to the indices in $\mathcal{I}$; for two index sets $\mathcal{I}$, $\mathcal{J}$, and a $M \times M$ dimensional matrix X, we use $X_{\mathcal{I} \times \mathcal{J}}$ 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 rate λ; we use $\mathrm{Ber}(\lambda)$ to denote a Bernoulli random variable with success probability λ; and we use $\mathrm{Mul}(x; \lambda)$ to denote a multinomial distribution over M outcomes with λ number of trials and event probability vector $x \in \mathbb{R}_+^M$ such that $\sum_i x_i = 1$.

1.2 Main Results

1.2.1 Recovering Rank-2 Probability Matrices

Throughout, we focus on a class of well-separated model parameters. The separation assumptions guarantee that the rank 2 matrix $\mathbb{B}$ is well-conditioned. Furthermore, this assumption also has natural interpretations in each of the different applications (to that of community detection, topic modeling, and HMMs).

All of our order notations are with respect to the vocabulary size M, which is asymptotically large. Also, we say that a statement is true "with high probability" if the failure probability of the statement is inverse poly in M; and we say a statement is true "with large probability" if the failure probability is of some small constant δ, which can be easily boosted to very smaller probabilities with repetitions.

Assumption 1 ($\Omega(1)$ separation)**.** *Assume that W is symmetric, where $w_L = w_R = w$ (all our results extend to the asymmetric case). Define the* marginal probability *vector, ρ and the* dictionary separation *vector as:*

$$\rho_i := \sum_k \mathbb{B}_{i,k}, \quad \Delta := w(p - q)\,. \tag{2}$$

Assume that the ℓ_1-norm of the dictionary separation is lower bounded by some constant $C_\Delta = \Omega(1)$,

$$\|\Delta\|_1 \geq C_\Delta. \tag{3}$$

Note that while in general the dictionary matrix D and the mixing matrix W are not uniquely identifiable from $\mathbb{B}$, there does exist an identifiable decomposition. Observe that the matrix $\mathrm{Cov}(\mathbb{B})$ admits a unique rank-1 decomposition: $\mathrm{Cov}(\mathbb{B}) := \mathbb{B} - \rho\rho^\top = \Delta\Delta^\top$, which also implies that:

$$\mathbb{B} = \rho\rho^\top + \Delta\Delta^\top. \tag{4}$$

It is this unique decomposition we seek to estimate, along with the matrix $\mathbb{B}$.

Now we are ready to state our main theorem.

Theorem 1.1 (Main theorem)**.** *Suppose we have access to N i.i.d. samples generated according to the rank 2 probability matrix $\mathbb{B}$ with structure given by* (1) *and satisfying the separation Assumption 1. For $\epsilon > 0$, with $N = \Theta(M/\epsilon^2)$ samples, our algorithm runs in time poly(M) and returns estimators $\widehat{B}$, $\widehat{\rho}$, $\widehat{\Delta}$, such that with large probability:*

$$\|\widehat{B} - \mathbb{B}\|_1 \leq \epsilon, \quad \|\widehat{\rho} - \rho\|_1 \leq \epsilon, \quad \|\widehat{\Delta} - \Delta\|_1 \leq \epsilon.$$

(here, the ℓ_1-norm of an $M \times M$ matrix P is simply defined as $\|P\|_1 = \sum_{i,j} |P_{i,j}|$).

First, note that we do not bound the spectral error $\|\widehat{B} - \mathbb{B}\|_2$ since when the marginal ρ is not roughly uniform, error bounds in terms of spectral distance are not particularly strong. A natural

error measure of estimation for probability distributions is total variation distance (equivalent to our ℓ_1 norm here). Second, note that naively estimating a distribution over M^2 outcomes requires order M^2 samples. Importantly, our algorithm utilizes the rank 2 structure of the underlying matrix $\mathbb{B}$ to achieve a sample complexity which is *precisely* linear in the vocabulary size M (without any additional log factors). The proof of the main theorem draws upon recent advances in the concentration of sparse random matrices from the community detection literature; the well characterized problem in the community detection literature can be viewed as a simple and special case of our problem, where the marginals are homogeneous (which we discuss later).

We now turn to the implications of this theorem to testing and learning problems.

1.2.2 Topic Models and Hidden Markov Models

One of the main motivations for considering the specific rank 2 structure on the underlying matrix $\mathbb{B}$ is that this structure encompasses the structure of the matrix of expected bigrams generated by both 2-topic models and two state HMMs. We now make these connections explicit.

Definition 1.2. *A 2-*topic model *over a vocabulary of size M is defined by a pair of distributions, p and q supported over M words, and a pair of topic* mixing weights *π_p and $\pi_q = 1 - \pi_p$. The process of drawing a bigram (i, j) consists of first randomly picking one of the two "topics" according to the mixing weights, and then drawing two independent words from the word distribution corresponding to the chosen topic. Thus the probability of seeing bigram (i, j) is $(\pi_p p_i p_j + \pi_q q_i q_j)$, and so the expected bigram matrix can be written as $\mathbb{B} = DWD^\top$ with $D = [p, q]$, and $W = \begin{bmatrix} \pi_p & 0 \\ 0 & \pi_q \end{bmatrix}$.*

The following corollary shows that estimation is possible with sample size *linear* in M:

Corollary 1.3. *(Learning 2-topic models) Suppose we are in the 2-topic model setting. Assume that $\pi_p(1 - \pi_p)\|p - q\|_1 = \Omega(1)$. There exists an algorithm which, given $N = \Omega(M/\epsilon^2)$ bigrams, runs in time poly(M) and with large probability returns estimates $\widehat{\pi}_p, \widehat{p}, \widehat{q}$ such that*

$$|\widehat{\pi}_p - \pi_p| < \epsilon,\ \|\widehat{p} - p\|_1 \leq \epsilon,\ \|\widehat{q} - q\|_1 \leq \epsilon.$$

Definition 1.4. *A* hidden Markov model *with 2 hidden states (s_p, s_q) and a size M observation vocabulary is defined by a 2×2 transition matrix T for the 2 hidden states, and two distributions of observations, p and q, corresponding to the 2 states.*

A sequence of N observations is sampled as follows: First, select an initial state according to the stationary distribution of the underlying Markov chain $[\pi_p, \pi_q]$; Then evolve the Markov chain according to the transition matrix T for N steps; For each $n \in \{1, \ldots, N\}$, the n-th observation in the sequence is generated by making an independent draw from either distribution p or q according to whether the Markov chain is in state s_p or s_q at the n-th timestep.

The probability that seeing a bigram (i, j) for the n and the $(n + 1)$-th observation is given by $\pi_p p_i(T_{p,p}p_j + T_{p,q}q_j) + \pi_q q_i(T_{q,p}p_j + T_{q,q}q_j)$, and hence the expected bigram matrix can be written as $\mathbb{B} = DWD^\top$ with $D = [p, q]$, and $W = \begin{bmatrix} \pi_p & 0 \\ 0 & \pi_q \end{bmatrix} \begin{bmatrix} T_{p,p} & 1 - T_{p,p} \\ 1 - T_{q,q} & T_{q,q} \end{bmatrix}$.

We have the following learning result:

Corollary 1.5. *(Learning 2-state HMMs) Suppose we are in the 2-state HMM setting. Assume that $\|p - q\|_1 \geq C_1$ and that π_p, $T_{p,p}$, $T_{q,q}$ are lower bounded by C_2 and upper bounded by $1 - C_2$, where both C_1 and C_2 are $\Omega(1)$. There exists an algorithm which, given a sampled chain of length $N = \Omega(M/\epsilon^2)$, runs in time poly(M) and returns estimates $\widehat{\pi}_p, \widehat{T}, \widehat{p}, \widehat{q}$ such that, with high probability, we have (that there is exists a permutation of the model such that)*

$$|\widehat{\pi}_p - \pi_p| < \epsilon, |\widehat{T}_{p,p} - T_{p,p}| < \epsilon, |\widehat{T}_{q,q} - T_{q,q}| < \epsilon,\ \|\widehat{p} - p\|_1 \leq \epsilon,\ \|\widehat{q} - q\|_1 \leq \epsilon.$$

Furthermore, it is sufficient for this algorithm to only utilize $\Omega(M/\epsilon^2)$ random bigrams and only $\Omega(1/\epsilon^2)$ random trigrams from this chain.

It is worth noting that the matrix of bigram probabilities does not uniquely determine the underlying HMM. However, one can recover the model parameters using sampled trigram sequences; this last step is straightforward (and sample efficient as it uses only an additional $\Omega(1/\epsilon^2)$ trigrams) when given an accurate estimate of $\mathbb{B}$ (see [6] for the moment structure in the trigrams).

1.2.3 Testing vs. Learning

The above theorem and corollaries are tight in an extremely strong sense. Both for the topic model and HMM settings, while we can learn the models using $\Omega(M)$ samples/observations, in both settings, it is information theoretically impossible to perform even the most basic property tests using fewer than $\Theta(M)$ samples.

In the case of 2-topic models, the community detection lower bounds [41][32][52] imply that $\Theta(M)$ bigrams are necessary to even distinguish between the case that the underlying model is simply the uniform distribution over bigrams versus the case of a 2-topic model in which each topic corresponds to a uniform distribution over disjoint subsets of $M/2$ words. We prove a stronger lower bound for the case of HMMs with two states, where we permit an estimator to have more information, namely the *full sequence* of observations (not merely bigram and trigram counts). Perhaps surprisingly, even with this extra information, we have the following lower bound:

Theorem 1.6. *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, $p = q = Uniform[M]$ from the case that $||p - q||_1 \geq 1/2$ using a sequence of fewer than $\Theta(M)$ observations.*

The proof of this theorem is given in Appendix 5.2, and amounts to a careful comparison of the processes of generating a uniformly random *path* in a graph, versus generating a path corresponding to a 2-state HMM for which there is significant correlation between consecutive observations. As an immediate corollary of this theorem, it follows that many natural properties of HMMs cannot be estimated using a sublinear length sequence of observations.

Corollary 1.7. *For HMMs with 2 states and emission distributions supported on a domain of size at most M, to estimate the* entropy rate *up to an additive constant $c \leq 1$ requires a sequence of $\Omega(M)$ observations.*

These strong lower bounds for property testing and estimation of HMMs are striking for several reasons. First, the core of our learning algorithm is a matrix reconstruction step that uses only the set of bigram counts (though we do use trigram counts for the final parameter recovery). Conceivably, one could benefit significantly from considering longer sequences of observations — even in HMMs that mix in constant time, there are detectable correlations between observations separated by $O(\log M)$ steps. Regardless, our lower bound shows that this is not the case. 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 brittleness of sublinear property testing and estimation as we deviate from the standard (unstructured) i.i.d sampling setting. In particular, while it may be natural to expect that testing and estimation would become rapidly more difficult as the number of hidden states of an HMM increase, we see here a (super-constant) increase in the difficulty of testing and estimation problems between the one state setting to the two state setting.

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* [27]. 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 in the form of Equation 1 as $\mathbb{B} = DWD^\top$ where $D = [p, q]$ 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 the past four years a number of transformations of the adjacency matrix have been proposed, after which the spectral approach will enable constant factor optimal detection (the problem of distinguishing the community setting from $\mathcal{G}(n, p)$ and reconstruction in this constant degree regime (see e.g. [22, 40, 32, 33]). In the past year, for both the *exact* recovery problem and the detection problem, the exact tradeoffs between α, β, and M were established, down to subconstant factors [41, 1, 36]. 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. [19, 2]).

Word Embeddings. On the more applied side, some of the most impactful advances in natural language processing over the past two years has been work on "word embeddings" [37, 35, 46, 9]. 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, jth entry corresponding to the total number of times the ith and jth word occur next to (or near) each other in a large corpus of text (e.g. wikipedia). 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

[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.

[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.

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 [45])—this is especially noticeable in the relatively poor quality of the embeddings for relatively rare words. The recent theoretical work [10] 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 [29, 39, 17], recovering low-rank structure [8, 7, 14], and learning or clustering various structured distributions such as Gaussian mixture models [20, 51, 38, 13, 28, 31, 24] and latent dirichlet allocation (a very popular topic model) [5]. A number of these methods essentially can be phrased as solving an inverse moments problem, and the work in [6] 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 focussed 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.

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 [43, 26, 47, 49], support size [44, 47], distance between distributions [47, 49, 48], and various hypothesis tests, such as whether two distributions are very similar, versus significantly different [11, 42, 50, 15], 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. [3, 4, 50], with stronger information theoretic optimality guarantees. There has also been significant work 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 [16, 12, 30, 21].

2 Outline of our estimation algorithm

In this section, we sketch the outline of our algorithm and explain the intuition behind the key ideas.

Given N samples drawn according to the probability matrix $\mathbb{B}$. Let B_N denote the matrix of empirical counts, and let $\frac{1}{N}B_N$ denote the average. By the Poisson assumption on sample size, we have that $[B_N]_{i,j} \sim \mathrm{Poi}(N\mathbb{B}_{i,j})$.

First, note that it is straightforward to obtain an estimate $\widehat{\rho}$ which is close to the true marginal ρ with ϵ accuracy in ℓ_1 norm with sample complexity $N = \Omega(M)$. Also, recall that $\mathbb{B} - \rho\rho^\top = \Delta\Delta^\top$ as per (4), hence after subtracting off the (relatively accurate) rank 1 matrix of marginals, we are essentially left with a rank 1 matrix recovery problem. Our algorithm seeks to accurately perform this rank-1 decomposition using a linear sample size, $N = \Theta(M)$.

Before introducing our algorithm, let us consider the naive approach of estimating Δ by taking the rank-1 truncated SVD of the matrix $(\frac{1}{N}B_N - \widehat{\rho}\widehat{\rho}^\top)$, which concentrates to $\Delta\Delta^\top$ in spectral distance asymptotically. Unfortunately, this approach leads to a sample complexity as large as $\Theta(M^2 \log M)$. In the linear sample size regime, the empirical counts matrix is a poor representation of the underlying distribution. Intuitively, due to the high sampling noise, the rows and columns of B_N 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. This

observation leads to the idea of pre-scaling the matrix so that every word (i.e. row/column) is roughly of unit variance. Indeed, with a slight modification of the truncated SVD, we can improve the sample complexity of this approach to $\Theta(M \log(M))$, which is nearly linear. It is, however, not obvious how to further improve this. Appendix E provides a detailed analysis of these aforementioned truncated SVD approaches.

Next, we briefly discuss the important ideas of our algorithm that lead to the linear sample complexity. Our algorithm consists of two phases: For a small constant ϵ_0, Phase I of our algorithm returns estimates $\widehat{\rho}$ and $\widehat{\Delta}$ both up to ϵ_0 accuracy in ℓ_1 norm with $\Theta(M)$ samples; After the first phase gets us off the ground, Phase II builds upon the output of Phase I to refine the estimation to any target accuracy ϵ with $\Theta(M/\epsilon^2)$ samples. The outline of the two phases are given in Algorithm 1 and Algorithm 2, separately, and the detailed analysis of the algorithms are deferred to Section 3 and 4.

The guarantees of the main algorithm follows immediately from Theorem 2.1 and 2.2.

Phase I: "binning" and "regularization" In Section 1, we drew the connection between our problem and the community detection problem in sparse random graphs. Recall that when the word marginals are roughly uniform, namely all in the order of $O(\frac{1}{M})$, the linear sample regime corresponds to the stochastic block model setup where the expected row / column sums are all in the order of $d_0 = \frac{N}{M} = \Omega(1)$. It is well-known that in this sparse regime, the adjacency matrix, or the empirical count matrix B_N in our problem, does not concentrate to the expectation matrix in the spectral distance. Due to some heavy rows with row sum in the order of $\Omega(\frac{\log M}{\log \log M})$, the leading eigenvectors are polluted by the local properties of these heavy nodes and do not reveal the global structure of the graph, which are precisely the desired information in expectation.

In order to enforce spectral concentration in the linear sample size regime, one of the many techniques is to tame the heavy rows and columns by setting them to 0. This simple idea was first introduced by [23], and followed by analysis works in [22] and many others. Recently in [33] and [34] the authors provided clean and clever proofs to show that *any* such "regularization" essentially leads to better spectral concentration for the adjacency matrix of random graphs whose row/column sums are roughly uniform in expectation.

Is it possible to leverage such "regularization" ideas in our problem where the marginal probabilities are not uniform? A natural candidate solution would be to partition the vocabulary $\mathcal{M}$ into bins of words according to the word marginals, so that the words in the same bin have roughly uniform marginals. Restricting our attention to the diagonal blocks of $\mathbb{B}$ whose indices are in the same bin, the expected row / column sums are indeed roughly uniform. Then we can regularize each diagonal block separately to guarantee spectral concentration, to which truncated SVD should then apply. Figure 1a visualizes the two operations of "binning" and "regularization".

Even though the "binning" idea seems natural, there are three concerns one needs to rigorously address in order to implement the idea:

1. We do not have access to the exact marginal ρ. With linear sample size, we only can estimate ρ up to constant accuracy in ℓ_1 norm. If we implement binning according to the empirical marginals, there is considerable probability that words with large marginals are placed in a bin intended for words with small marginals — which we call "spillover effect". When directly applied to the empirical bins with such spillover, the existing results of "regularization" in [34] do not lead to the desired concentration result.

2. When restricting to each diagonal block corresponding to a bin, we are throwing away all the sample counts outside the block. This greatly reduces the effective sample size, and it is not obvious that we retain enough samples in each diagonal block to guarantee meaningful concentration results and subsequent estimation. This is particularly worrying because we may use a super-constant number of bins, and hence throw away all but a subconstant fraction of data for some words.

Input: $2N$ sample counts.

Output: Estimator $\widehat{\rho}$, $\widehat{\Delta}$, $\widehat{B}$.

Divide the sample counts into two independent batches of equal size N, and construct two empirical count matrices B_{N1} and B_{N2}.

1. (**Estimate the marginal**) $\widehat{\rho}_i = \frac{1}{N}\sum_{j\in[M]}[B_{N1}]_{i,j}$.

2. (**Binning**) Partition the vocabulary $\mathcal{M}$ into:
   $\widehat{\mathcal{I}}_0 = \left\{i : \widehat{\rho}_i < \frac{\epsilon_0}{M}\right\}$, $\widehat{\mathcal{I}}_{\log} = \left\{i : \widehat{\rho}_i > \frac{\log(M)}{M}\right\}$, $\widehat{\mathcal{I}}_k = \left\{i : \frac{e^k}{M} \le \widehat{\rho}_i \le \frac{e^{k+1}}{M}\right\}$, $k = 1 : \log\log(M)$.

3. (**Estimate the separation $\widehat{\Delta}$**)

   (a) (**Estimate $\widehat{\Delta}_{\widehat{\mathcal{I}}_{\log}}$**) (up to sign flip)

   If $\sum \widehat{\rho}_{\widehat{\mathcal{I}}_{\log}} < \epsilon_0$, set $\widehat{\Delta}_{\widehat{\mathcal{I}}_{\log}} = 0$, else

   i. (**Rescaling**): Set $E = diag(\widehat{\rho}_{\widehat{\mathcal{I}}_{\log}})^{-1/2}[B_{N2} - \widehat{\rho}\widehat{\rho}^\top]_{\widehat{\mathcal{I}}_{\log}\times\widehat{\mathcal{I}}_{\log}} diag(\widehat{\rho}_{\widehat{\mathcal{I}}_{\log}})^{-1/2}$.

   ii. (**t-SVD**): Let $u_{\log}u_{\log}^\top$ be the rank-1 truncated SVD of E.

   iii. Set $v_{\log} = diag(\widehat{\rho}_{\widehat{\mathcal{I}}_{\log}})^{1/2}u_{\log}$.

   (b) (**Estimate $\widehat{\Delta}_{\widehat{\mathcal{I}}_k}$**) (up to sign flip)

   If $\sum \widehat{\rho}_{\widehat{\mathcal{I}}_k} < \epsilon_0 e^{-k}$, set $\widehat{\Delta}_{\widehat{\mathcal{I}}_k} = 0$, else

   i. (**Regularization**): Set $\widetilde{B} = [B_{N2}]_{\widehat{\mathcal{I}}_k\times\widehat{\mathcal{I}}_k}$, set $d_k^{\max} = N(\sum \widehat{\rho}_{\widehat{\mathcal{I}}_k})\frac{e^{k+\tau}}{M}$.
   If a row/column of $\widetilde{B}$ has sum larger than $2d_k^{max}$, set the entire row/column to 0.
   Set $E = (\widetilde{B} - \widehat{\rho}_{\widehat{\mathcal{I}}_k}\widehat{\rho}_{\widehat{\mathcal{I}}_k}^\top)$.

   ii. (**t-SVD**): Let $v_k v_k^\top$ be the rank-1 truncated SVD of E.

   (c) (**Estimate $\widehat{\Delta}_{\widehat{\mathcal{I}}_0}$**) Set $\widehat{\Delta}_{\widehat{\mathcal{I}}_0} = 0$.

   (d) (**Stitching $\widehat{\Delta}_{\widehat{\mathcal{I}}_k}$'s**) Fix $k^* = \arg\max_k \|v_k\|$, set $\widehat{\Delta}_{\widehat{\mathcal{I}}_k^*} = v_{k^*}$.

   For all $k \neq k^*$, consider the block $E_{k,k^*} = [B_{N2} - \widehat{\rho}\widehat{\rho}^\top]_{\widehat{\mathcal{I}}_k\times\widehat{\mathcal{I}}_{k^*}}$. Set $\alpha_k = \sum_{i,j}\frac{[E_{k,k^*}]_{i,j}}{(v_k)_i(v_k^*)_j}$;

   Set $\widehat{\Delta}_{\widehat{\mathcal{I}}_k} = (\mathbf{1}[\alpha_k > 0] - \mathbf{1}[\alpha_k < 0])v_k$.

4. **Return** $\widehat{\rho}$, $\widehat{\Delta}$, and $\widehat{B} = \widehat{\rho}\widehat{\rho}^\top + \widehat{\Delta}\widehat{\Delta}^\top$.

Algorithm 1: Phase I

3. 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, which includes the off-diagonal blocks.

Phase I (Algorithm 1) capitalizes on these natural ideas of "binning" and "regularization", and avoids the above potential pitfalls. We show that the algorithm satisfies the following guarantees:

Theorem 2.1 (Main theorem 1. $\Theta(M)$ sample complexity for achieving constant accuracy in ℓ_1 norm)**.** *Fix ϵ_0 to be a small constant. Given $N = \Theta(M)$ samples, with large probability, Phase I*

Input: Estimator $\widehat{\rho}$ and $\widehat{\Delta}$ from Phase I. N sample counts.

Output: Refined estimator $\widehat{\rho}$, $\widehat{\Delta}$, $\widehat{B}$.

1. (**Construct anchor partition**)

   Set $\mathcal{A} = \phi$. For $k = 1, \ldots, \log\log M, \log$:

   If $\|\widehat{\Delta}_{\widehat{\mathcal{I}}_k}\|_2 \le (\frac{\sqrt{d_k^{\max}}}{N})^{1/2}$, skip the bin, else, set $\mathcal{A} = \mathcal{A} \cup \{i \in \widehat{\mathcal{I}}_k : \widehat{\Delta}_i > 0\}$.

2. (**Estimate anchor matrix**)

   Set $B_{\mathcal{A}} = \begin{bmatrix} \sum_{i\in\mathcal{A}, j\in\mathcal{A}}[B_N]_{i,j} & \sum_{i\in\mathcal{A}, j\in\mathcal{A}^c}[B_N]_{i,j} \\ \sum_{i\in\mathcal{A}^c, j\in\mathcal{A}}[B_N]_{i,j} & \sum_{i\in\mathcal{A}^c, j\in\mathcal{A}^c}[B_N]_{i,j} \end{bmatrix}$. Set vector $b = \begin{bmatrix} \sum_{i\in\mathcal{A}, j\in\mathcal{M}}[B_N]_{i,j} \\ \sum_{i\in\mathcal{A}^c, j\in\mathcal{M}}[B_N]_{i,j} \end{bmatrix}$.

   Set $aa^\top$ to be rank-1 truncated SVD of the 2×2 matrix $(B_{\mathcal{A}} - bb^\top)$.

3. (**Refine the estimation:**)

   Set $\begin{bmatrix} \widehat{\rho}^\top \\ \widehat{\Delta}^\top \end{bmatrix} = [a, b]^{-1} \begin{bmatrix} \sum_{i\in\mathcal{A}}[B_N]_{i,\mathcal{M}} \\ \sum_{i\in\mathcal{A}^c}[B_N]_{i\mathcal{M}} \end{bmatrix}$

4. (**Return**) $\widehat{\rho}$, $\widehat{\Delta}$ and $\widehat{B} = \widehat{\rho}\widehat{\rho}^\top + \widehat{\Delta}\widehat{\Delta}^\top$.

Algorithm 2: Phase II

(Algorithm 1) estimates ρ and Δ up to ϵ_0 accuracy in ℓ_1 norm, namely

$$\|\widehat{\rho} - \rho\|_1 < \epsilon_0, \quad \|\widehat{\Delta} - \Delta\|_1 < \epsilon_0, \quad \|\widehat{B} - \mathbb{B}\|_1 < \epsilon_0.$$

Phase II: "Anchor partition" Under the Assumption of $\Omega(1)$ separation, Phase II of our algorithm makes use of the estimates of Δ computed by Phase I, to refine the estimates of ρ and Δ.

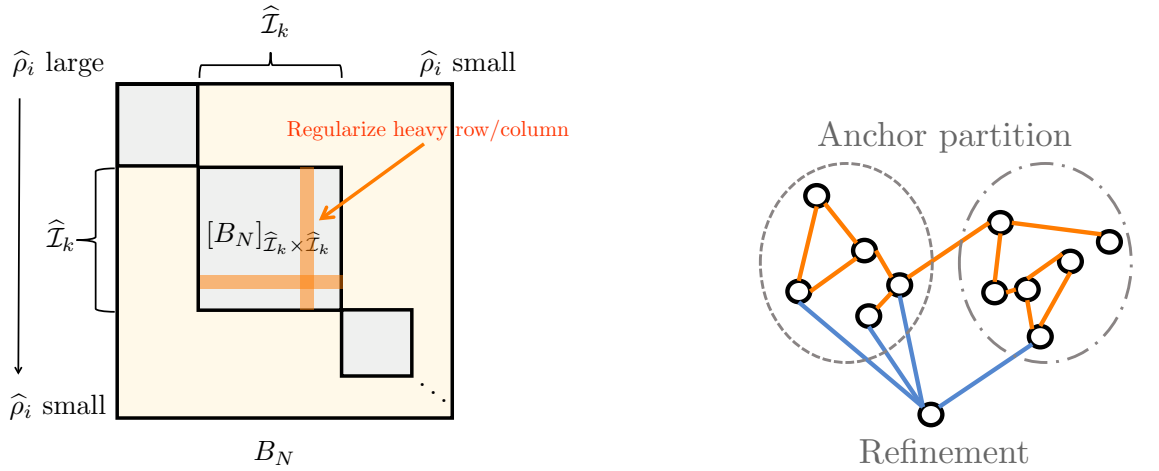
The key to this refining process is to construct an "anchor partition", which is a bi-partition of the vocabulary $\mathcal{M}$ based on the signs of the estimate of separation $\widehat{\Delta}$ given by Phase I. We collapse the $M \times M$ matrix B_N into a 2×2 matrix corresponding to the bi-partition, and accurately estimate the 2×2 matrix with the N samples. Given this extremely accurate estimate of this 2×2 anchor matrix, we can now iteratively refine our estimates of ρ_i and Δ_i for each word i by solving a simple least square fitting problem.

The above description may seem opaque, but similar ideas — estimation refinement based on some crude global information — has appeared in many works for different problems. For example, in a recent paper [19] on community detection, after obtaining a crude classification of nodes using spectral algorithm, one round of a "correction" routine is applied to each node based on its connections to the graph partition given by the first round. This refinement immediately leads to an optimal rate of recovery. Another example is given in [18] in the context of solving random quadratic equations, where local refinement of the solution follows the spectral method initialization. Figure 1b visualize the example of community detection. In our problem, the nodes are the M words, the edges are the sample counts, and instead of re-assigning the label to each node in the refinement routine, we re-estimate the ρ_i and Δ_i for each word i.

Theorem 2.2 (Main theorem 2. $\Theta(M/\epsilon^2)$ sample complexity to achieve ϵ accuracy in ℓ_1 norm)**.** *Assume that $\mathbb{B}$ satisfies the $\Omega(1)$ separation assumption. Given N samples, with large probability,*

Phase II of our algorithm (Algorithm 2) estimates ρ and Δ up to accuracy in ℓ_1 norm:

$$\|\widehat{\rho} - \rho\|_1 < O(\sqrt{\frac{M}{N}}), \quad \|\widehat{\Delta} - \Delta\|_1 < O(\sqrt{\frac{M}{N}}), \quad \|\widehat{B} - \mathbb{B}\|_1 < O(\sqrt{\frac{M}{N}}).$$



(a) Binning and regularization

(b) Anchor partition and refinement

Figure 1: The key algorithmic ideas of our algorithm.

3 Algorithm Phase I, achieving constant ϵ_0 accuracy

In this section, we outline the proof for Theorem 2.1, the detailed proofs are provided in Section A and B in the appendix.

We denote the ratio between sample size and the vocabulary size by

$$d_0 = \frac{N}{M}. \tag{5}$$

Throughout our discussion for Phase I algorithm, we assume that $d_0 = \Theta(1)$ to be some fixed large constant.

Given N samples, the goal is to estimate the word marginal vector ρ as well as the dictionary separation vector Δ up to constant accuracy in ℓ_1 norm. We denote the estimates by $\widehat{\rho}$ and $\widehat{\Delta}$. Also, we estimate the underlying probability matrix $\mathbb{B}$ with

$$\widehat{B} = \widehat{\rho}\widehat{\rho}^\top + \widehat{\Delta}\widehat{\Delta}^\top.$$

Note that since $\|\Delta\|_1 \le \|\rho\|_1 = 1$, constant ℓ_1 norm accuracy in $\widehat{\rho}$ and $\widehat{\Delta}$ immediately lead to constant accuracy of $\widehat{B}$ also in ℓ_1 norm.

First, we show that it is easy to estimate the marginal probability vector ρ up to constant accuracy.

Lemma 3.1 (Estimate the word marginal probability ρ)**.** *Given the empirical count matrix B_{N1} constructed with the first batch of N sample counts, consider the estimate of the marginal probabilities:*

$$\widehat{\rho}_i = \frac{1}{N} \sum_{j \in \mathcal{M}} [B_{N1}]_{i,j}. \tag{6}$$

With large probability, we can bound the estimation accuracy by:

$$\|\widehat{\rho} - \rho\|_1 \le O(\sqrt{\frac{1}{d_0}}). \tag{7}$$

The hard part is to estimate the separation vector Δ up to constant accuracy in ℓ_1 norm with linear number of sample counts, namely $d_0 = \Theta(1)$. Recall that naively taking the rank-1 truncated SVD of $(\frac{1}{N}B_N - \widetilde{\rho}\widehat{\rho}^\top)$ fails to reveal any information about $\Delta\Delta^\top$, since in the linear sample size regime, the leading eigenvectors of B_N are mostly dominated by the statistical noise of the words with large marginal probabilities. Our Phase I algorithm achieves this with more delicate steps. We analyze each step in the next 5 subsections, structured as follows:

1. In Section 3.1, we introduce the binning argument and the necessary notations for the rest of the section. We bin the vocabulary $\mathcal{M}$ according to the estimates of word marginals, i.e. $\widehat{\rho}$, and we call a bin heavy or light according to the typical word marginals in that bin.

2. In Section 3.2 we analyze how to estimate the entries of Δ restricted to the heaviest bin (up to some common sign flip). Because the marginal probabilities of words in this bin are sufficiently large, truncated SVD can be applied to the properly scaled diagonal block of the empirical average count matrix.

3. In Section 3.3 we analyze how to estimate the entries of Δ restricted to all the other bins (up to some common sign flip), by examining the corresponding diagonal blocks in the empirical average count matrix.

   The main challenge here is that due to the estimation error of the word marginal vector $\widehat{\rho}$, the binning is not perfect, in the sense that a lighter bin may include some heavy words by chance. Lemma 3.4 is the key technical lemma, which shows that with high probability, such spillover effect is very small *for all bins* with high probability. Then we leverage the clever proof techniques from [34] to show that if spillover effect is small, regularized truncated SVD can be applied to estimate the entries of Δ restricted to each bin.

4. In Section 3.4, we analyze the lightest bin.

5. In Section 3.5 we show how to fix the sign flips across different bins, by using the off-diagonal blocks, so that we can concatenate different sections of $\widehat{\Delta}$ to obtain an estimator for the entire separation vector Δ.

3.1 Binning according to empirical marginal distribution

Instead of tackling the empirical count matrix B_N as a whole, we focus on its diagonal blocks and analyze the spectral concentration restricted to each block separately. Since the entries $\mathbb{B}_{i,j}$ restricted to each diagonal block are roughly uniform, the block hopefully concentrates well, so that we can estimate segments of the separation vector Δ by using truncated SVD with the "regularization" trick.

For any set of words $\mathcal{I}$, we use $[B_N]_{\mathcal{I},\mathcal{I}}$ to denote the diagonal block of B_N whose indices are in the set $\mathcal{I}$. Note that when restricting to the diagonal block, the rank 2 decomposition in (4) is given by $\mathbb{B}_{\mathcal{I},\mathcal{I}} = \rho_{\mathcal{I}}\rho_{\mathcal{I}}^\top + \Delta_{\mathcal{I}}\Delta_{\mathcal{I}}^\top$.

Empirical binning We partition the vocabulary $\mathcal{M}$ according to the empirical marginal $\widehat{\rho}$ in (6):

$$\widehat{\mathcal{I}}_0 = \left\{i : \widehat{\rho}_i < \frac{\epsilon_0}{M}\right\}, \quad \widehat{\mathcal{I}}_k = \left\{i : \frac{e^{k-1}}{M} \le \widehat{\rho}_i \le \frac{e^k}{M}\right\}, \quad \widehat{\mathcal{I}}_{\log} = \left\{i : \widehat{\rho}_i > \frac{\log(M)}{M}\right\}.$$

We call this *empirical binning* to emphasize the dependence on the empirical estimator $\widehat{\rho}$, which is a random variable built from the first batch of N sample counts. We call $\widehat{\mathcal{I}}_0$ the *lightest empirical bin*, and $\widehat{\mathcal{I}}_{\log}$ the *heaviest empirical bin*, and $\widehat{\mathcal{I}}_k$ for $1 \le k \le \log\log M$ the *moderate empirical bins*.

For the analysis, we further define the *exact bins* according to the exact marginal probabilities:

$$\mathcal{I}_0 = \left\{i : \rho_i < \frac{\epsilon_0}{M}\right\}, \quad \mathcal{I}_k = \left\{i : \frac{e^{k-1}}{M} \le \rho_i \le \frac{e^k}{M}\right\}, \quad \mathcal{I}_{log} = \left\{i : \rho_i > \frac{\log(M)}{M}\right\}. \tag{8}$$

Spillover effect As N increases asymptotically, we have $\widehat{\mathcal{I}}_k$ coincides with $\mathcal{I}_k$ for each k. However, in the linear sample size regime, the estimation error in $\widehat{\rho}$ cause the following two *spillover effects*:

1. Words from the heavy bin $\mathcal{I}_{k'}$, for k' much larger than k, are placed in the empirical bin $\widehat{\mathcal{I}}_k$.
2. Words from the exact bin $\mathcal{I}_k$ escape from the corresponding empirical bin $\widehat{\mathcal{I}}_k$;

The hope, that we can have good spectral concentration in each diagonal block $[B_{N2}]_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$, crucially relies on the fact that the entries $\mathbb{B}_{i,j}$ restricted to this block are roughly uniform. However, the hope may be ruined by the spillover effects, especially the first one. In the following sections, we show that with high probability the spillover effects are small for all empirical bins of considerable probability mass, in particular:

1. The total marginal probability of the words in the empirical bin $\widehat{\mathcal{I}}_k$, that are from faraway exact bins, namely $\cup_{\{k':k'\geq k+\tau\}}\mathcal{I}_{k'}$, is small and in the order of $O(e^{-e^k d_0})$ (see Lemma 3.4).
2. Most words of $\mathcal{I}_k$ stays within the nearest empirical bins, namely $\cup_{\{k':k-\tau\leq k'\leq k+\tau\}}\widehat{\mathcal{I}}_{k'}$, (see Lemma 4.5).

Throughout the discussion, we fix some small constant number τ to be:

$$\tau = 1 \tag{9}$$

Notations To analyze the spillover effects, we define some additional quantities.

We define the total marginal probability mass in the empirical bins to be:

$$W_k = \sum_{i\in\widehat{\mathcal{I}}_k} \rho_i, \tag{10}$$

and let $M_k = |\widehat{\mathcal{I}}_k|$ denote the total number of words in the empirical bin. We also define $\widehat{W}_k = \sum_{i\in\widehat{\mathcal{I}}_k} \widehat{\rho}_i$.

We use $\widehat{\mathcal{J}}_k$ to denote the set of the words from much heavier bins that are spilled over into the empirical bin $\widehat{\mathcal{I}}_k$ (recall that τ is a small constant):

$$\widehat{\mathcal{J}}_k = \widehat{\mathcal{I}}_k \cap (\cup_{\{k':k'\geq k+\tau\}}\mathcal{I}_{k'}), \tag{11}$$

and let $\widehat{\mathcal{L}}_k$ denote the "good words" in the empirical bin $\widehat{\mathcal{I}}_k$:

$$\widehat{\mathcal{L}}_k = \widehat{\mathcal{I}}_k \backslash \widehat{\mathcal{J}}_k. \tag{12}$$

where τ We also denote the total marginal probability mass of the heavy spillover words $\widehat{\mathcal{J}}_k$ by:

$$\overline{W}_k = \sum_{i\in\widehat{\mathcal{J}}_k} \rho_i. \tag{13}$$

Note that these quantities are random variables determined by the randomness of the first batch of N samples, via the empirical binning. These are fixed quantities when we consider the empirical count matrix with the second batch of N samples.

Define the upper bound of the "typical" word marginal in the k-th empirical bin to be:

$$\overline{\rho}_k = e^{k+\tau}/M,$$

and let $d_k^{\max}$ denote the expected max row/column sum of the diagonal block corresponding the k-th bin:

$$d_k^{\max} = NW_k\overline{\rho}_k. \tag{14}$$

3.2 Estimate Δ restricted to the heaviest empirical bin

First, we show that the empirical marginal probabilities restricting the heaviest bin concentrate much better than what Lemma 3.1 implies.

Lemma 3.2 (Concentration of marginal probabilities in the heaviest bin)**.** *With high probability, for all the words with marginal probability $\rho_i \geq \log(M)/M$, we have that for some universal constant C_1, C_2,*

$$C_1 \leq \widehat{\rho}_i/\rho_i \leq C_2. \tag{15}$$

Lemma 3.2 says that with high probability, we can estimate the marginal probabilities for every words in the heaviest bin with constant multiplicative accuracy. Note that it also suggests that actually we do not need to worry about the spillover effect of the words from $\mathcal{I}_{\log}$ placed in much lighter bins, since with high probability, all the words stay in the bin $\widehat{\mathcal{I}}_{\log}$ and a constant number of adjacent empirical bins.

Next two lemmas show that with square root re-scaling, truncated SVD works to estimate the segment of separation restricted to the empirical heaviest bin.

Lemma 3.3 (Estimate Δ restricted to the heaviest empirical bin)**.** *Suppose that $\widehat{W}_{\log} = \sum \widehat{\rho}_{\widehat{\mathcal{I}}_{\log}} > \epsilon_0$. Define $\widehat{D}_{\widehat{\mathcal{I}}_{\log}} = diag(\widehat{\rho}_{\widehat{\mathcal{I}}_{\log}})$, and consider the diagonal block corresponding to the heaviest empirical bin $\widehat{\mathcal{I}}_{\log}$:*

$$E = \widehat{D}_{\widehat{\mathcal{I}}_{\log}}^{-1/2}\left(\frac{1}{N}[B_{N2}]_{\widehat{\mathcal{I}}_{\log},\widehat{\mathcal{I}}_{\log}} - \widehat{\rho}_{\widehat{\mathcal{I}}_{\log}}\widehat{\rho}_{\widehat{\mathcal{I}}_{\log}}^{\top}\right)\widehat{D}_{\widehat{\mathcal{I}}_{\log}}^{-1/2}. \tag{16}$$

Let $uu^{\top}$ denote the rank 1 truncated SVD of matrix E, set $v_{\log} = \widehat{D}_{\widehat{\mathcal{I}}_{\log}}^{1/2} u$. With high probability over the second batch of N samples, we can estimate $\Delta_{\widehat{\mathcal{I}}_{\log}}$, the dictionary separation vector restricted to the heaviest empirical bin, with $v_{\log}$ up to sign flip with accuracy:

$$\min\{\|\Delta_{\widehat{\mathcal{I}}_{\log}} - v_{\log}\|_1, \|\Delta_{\widehat{\mathcal{I}}_{\log}} + v_{\log}\|_1\} = O\left(\min\left\{\frac{1/d_0^{1/2}}{\|\Delta_{\widehat{\mathcal{I}}_{\log}}\|_1}, \quad 1/d_0^{1/4}\right\}\right). \tag{17}$$

The two cases in the above bound correspond to whether the separation is large or small, compared to the statistical noise from sampling, which is in the order $1/d_0^{1/4}$. If the bin contains a large separation, then the bound follows the standard Wedin's perturbation bound; if the separation is small, i.e. $\|\Delta_{\widehat{\mathcal{I}}_{\log}}\|_1 \ll 1/d_0^{1/4}$, then the bound $1/d_0^{1/4}$ just corresponds to the magnitude of the statistical noise.

3.3 Estimate Δ restricted to the $\log\log(M)$ moderate empirical bins

In this section, we show that the spillover effects are small for all the moderate bins (Lemma 3.4). In particular, we upper bound the total spillover marginal $\overline{W}_k$ *for all* k with high probability. Provided that the spillover effects are small, we show that (Lemma 3.5 and Lemma 3.6) truncated SVD with regularization can be applied to each diagonal block of the empirical count matrix B_{N2}, to estimate the entries of the separation vector Δ restricted to each bin. The proofs of this section are provided in Section B in the appendix.

Lemma 3.4 (With high probability, spillover from much heavier bins is small)**.** *With* high *probability over the first batch of N sample counts, for all empirical bins $\{\widehat{\mathcal{I}}_0, \widehat{\mathcal{I}}_1, \ldots \widehat{\mathcal{I}}_{\log\log(M)}\}$, we can bound the total marginal probability of spillover from much heavier bins, i.e. $\overline{W}_k$ defined in* (13)*, by:*

$$\overline{W}_k \leq 2e^{-e^{\tau+k}d_0/2}. \tag{18}$$

Also, if $W_k > \epsilon_0 e^{-k}$, we can bound the number of spillover words $\overline{M}_k$ by:

$$\overline{M}_k \leq M_k / d_k^{\max}. \tag{19}$$

Recall that τ is set in (9)*, W_k is defined in* (10)*, and $d_k^{\max} = NW_k\overline{\rho}_k$ is defined in* (14)*.*

Now consider $[B_{N2}]_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$, the diagonal block corresponding to the empirical bin $\widehat{\mathcal{I}}_k$. To ensure the spectral concentration of this diagonal block, we "regularize" it by removing the rows and columns with very large sum. The spectral concentration of the block with the remaining elements leads to an estimate of the separation vector $\Delta_{\widehat{\mathcal{I}}_k}$ restricted to $\widehat{\mathcal{L}}_k$, the set of "good words" defined in (12). To make the operation of "regularization" more precise, we need to introduce some additional notations.

Define $\widetilde{\rho}_k$ to be a vector with the same length as $\rho_{\widehat{\mathcal{I}}_k}$, with the same entries for the good words, and set the entries corresponding to the spillover set $\widehat{\mathcal{J}}_k$ to be 0, namely

$$(\widetilde{\rho}_k)_i = \rho_i \mathbf{1}\left[i \in \widehat{\mathcal{L}}_k\right].$$

Similarly define vector $\widetilde{\Delta}_k$ to be the separation restricted to the good words in the empirical bins:

$$(\widetilde{\Delta}_k)_i = \Delta_i \mathbf{1}\left[i \in \widehat{\mathcal{L}}_k\right]. \tag{20}$$

We define the matrix $\widetilde{\mathbb{B}}_k$ (of the same size as $[B_{N2}]_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$):

$$\widetilde{\mathbb{B}}_k = \widetilde{\rho}_k \widetilde{\rho}_k^\top + \widetilde{\Delta}_k \widetilde{\Delta}_k^\top. \tag{21}$$

Recall that the expected max row sum of the diagonal block is given by $d_k^{\max} = NW_k\overline{\rho}_k$ defined in (14). Let $\widehat{\mathcal{R}}_k$ denote the indices of the rows and columns in $[B_{N2}]_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}_k}$ whose row sum or column sum are larger than $2d_k^{\max}$, namely

$$\widehat{\mathcal{R}}_k = \left\{ i \in \widehat{\mathcal{I}}_k : \sum_{j \in \widehat{\mathcal{I}}_k} [B_{N2}]_{i,j} > 2d_k^{\max} \text{ or } \sum_{j \in \widehat{\mathcal{I}}_k} [B_{N2}]_{j,i} > 2d_k^{\max} \right\}. \tag{22}$$

Starting with $\widetilde{B}_k = [B_{N2}]_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$, we set all the rows and columns of $\widetilde{B}_k$ indexed by $\widehat{\mathcal{R}}_k$ to 0.

Note that by definition the rows and columns in $\widetilde{B}_k$ and $\widetilde{\mathbb{B}}_k$ that are zero-ed out do not necessarily coincide. However, the next lemma shows that $\widetilde{B}_k$ concentrates to $\widetilde{\mathbb{B}}_k$ in the spectral distance.

Lemma 3.5 (Concentration of regularized diagonal block $\widetilde{B}_k$.)**.** *Suppose that the marginal of the bin $\widehat{\mathcal{I}}_k$ is large enough $W_k = \sum \rho_{\widehat{\mathcal{I}}_k} > \epsilon_0 e^{-k}$. With high probability at least $(1 - M_k^{-r})$, where r is some universal constant, we have*

$$\left\| \frac{1}{N} \widetilde{B}_k - \widetilde{\mathbb{B}}_k \right\|_2 \leq C r^{1.5} \frac{\sqrt{d_k^{\max}} \log^2 d_k^{\max}}{N}, \tag{23}$$

Proof. Detailed proof of this lemma is provided in Section B in the appendix. Here we highlight the key steps of the proof.

In Figure 2, the rows and the columns of $[B_N]_{\widehat{\mathcal{I}}_k \times \widehat{\mathcal{I}}_k}$ are sorted according to the exact marginal probabilities of the words in ascending order. The rows and columns that are set to 0 by regularization are shaded. Consider the block decomposition according to the good words $\widehat{\mathcal{L}}_k$ and the spillover words $\widehat{\mathcal{J}}_k$. We bound the spectral distance of the 4 blocks (A_1, A_2, A_3, A_4) separately. The bound for the entire matrix $\widehat{B}_k$ is then an immediate result of triangle inequality.

For block A_1 whose rows and columns all correspond to the "good words" with roughly uniform marginals, we show its concentration by applying the result in [34]. For block A_2 and A_3, we want to show that after regularization the spectral norm of these two blocks are small. Intuitively, the expected row sum and the expected column sum of block A_2 are bounded by $2d_k^{\max}$ and $2d_k^{\max}\frac{\overline{W}_k}{W_k} = O(1)$, as a result of the bound on the spillover words $\overline{W}_k$ in Lemma 3.4. Therefore the spectral norm of the block are likely to be bounded by $O(\sqrt{d_k^{\max}})$, which we show rigorously with high probability arguments. Lastly for block A_4, which rows and columns all correspond to the spillover words. We show that the spectral norm of this block is very small as a result of the small spillover marginal $\overline{W}_k$.

It is also important to note that given $W_k > \epsilon_0 e^{-k}$ and conditional on the high probability even that $\overline{W}_k \le 2e^{-e^k d_0}$, we can write $d_k^{\max}$ also as $d_k^{\max} = NM_k\overline{\rho}_k^2$. □

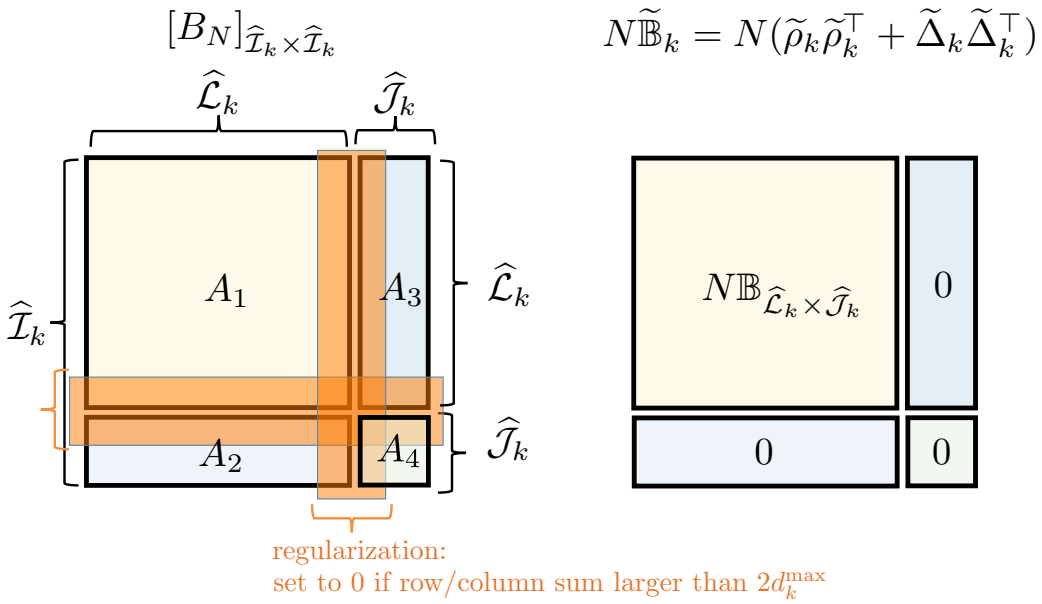


Figure 2: block decomposition of the diagonal block of B_{N2} corresponding to $\widehat{\mathcal{I}}_k$.

Lemma 3.6 (Given spectral concentration of block $\widetilde{B}_k$, estimate the separation $\widetilde{\Delta}_k$)**.** *Suppose that the marginal of the bin $\widehat{\mathcal{I}}_k$ is large enough $W_k = \sum \rho_{\widehat{\mathcal{I}}_k} > \epsilon_0 e^{-k}$. Let $v_k v_k^\top$ be the rank-1 truncated SVD of the regularized block $\left(\frac{1}{N}\widetilde{B}_k - \widehat{\rho}_{\widehat{\mathcal{I}}_k}\widehat{\rho}_{\widehat{\mathcal{I}}_k}^\top\right)$. With high probability over the second batch of N samples,*

$$\min\{\|\widetilde{\Delta}_k - v_k\|_2, \|\widetilde{\Delta}_k + v_k\|_2\} = O\left(\min\left\{\frac{\sqrt{d_k^{\max}}\log^2 d_k^{\max}}{N}\frac{1}{\|\Delta_{\widehat{\mathcal{I}}_k}\|_2}, \quad \left(\frac{\sqrt{d_k^{\max}}\log^2 d_k^{\max}}{N}\right)^{1/2}\right\}\right). \tag{24}$$

Namely v_k is an estimate of the vector $\widetilde{\Delta}_k$ (defined as in (21)), up to some unknown sign flip.

3.4 Estimate Δ restricted to the lightest empirical bin.

Claim 3.7 (Estimate separation restricted to the lightest bin)**.** *Setting $\widehat{\Delta}_{\widehat{\mathcal{I}}_0} = 0$ only incurs an ℓ_1 error of a small constant, and this is simply because the total marginal of words in the lightest bin with high probability can be bounded by a small constant:*

$$\|\Delta_{\widehat{\mathcal{I}}_0}\|_1 \le \|\rho_{\widehat{\mathcal{I}}_0}\|_1 \le \|\rho_{\mathcal{I}_0}\|_1 + \overline{W}_0 \le \frac{\epsilon_0}{M}M + e^{-d_0/2} = O(\epsilon_0),$$

where we used the assumption that $d_0 \geq 1/\epsilon_0^4$.

3.5 Stitching the segments of $\widehat{\Delta}$ to reconstruct an estimation of the matrix

Form 4 sets, S1 =words w in first bin with t_w ¿ 0, S1' = words w in first bin with $t_w <= 0$, S2 =words w in second bin with $t_w > 0$, S2' = words w in second bin with $t_w >= 0$. Merge all words in each of the sets, and compare $Pr[S1, S2]/Pr[S2]$ to $Pr[S1, S2']/Pr[S2']$. If the first quantity is larger, then t and t' have the same sign, otherwise negate t'.

Claim 3.8 (Pairwise comparison of bins to fix sign flips)**.** *Given* v_k *for all* k *as estimation for* $\Delta_{\widehat{\mathcal{I}}_k}$*'s up to sign flips. Fix* k^* *to be one good bin (with large bin marginal and large separation). For all other good bins* k*,*

examine the off-diagonal block $E_{k,k^*} = [\frac{1}{N} B_{N2} - \widehat{\rho}\widehat{\rho}^\top]_{\widehat{\mathcal{I}}_k, \widehat{\mathcal{I}}'_k}$*, which in expectation equals* $\Delta_k \Delta_{k^*}^\top = c_k v_k v_{k^*}^\top$*, thus we can estimate the sign flip of* k *with respect to* k^**, by*

$$\alpha_{kk^*} = \sum_{i,j} \frac{[E_{k,k^*}]_{i,j}}{(v_k)_i (v_{k^*})_j},$$

and set $v_k = (\mathbf{1}[\alpha_{kk'} > 0] - \mathbf{1}[\alpha_{kk'} < 0])v_k$*. Finally, we have* $\widehat{\Delta} = [\widehat{\Delta}_{\widehat{\mathcal{I}}_0}, \widehat{v}_1, \ldots, \widehat{v}_{\log\log M}, \widehat{v}_{\log}]$*, as an estimator for* Δ*.*

Finally, concatenate the segments of $\widehat{\Delta}$, we can summarize to bound the overall estimation error in ℓ_1 norm:

Lemma 3.9 (Accuracy of $\widehat{\Delta}$ of Phase I)**.** *For fixed* $\epsilon_0 = \Omega(1)$ *to be a small constant, if* $d_0 =: N/M \geq 1/\epsilon_0^4$*, with large probability, Phase I algorithm can estimate the separation vector* Δ *with constant accuracy in* ℓ_1 *norm:*

$$\|\widehat{\Delta} - \Delta\| = O(\epsilon_0).$$

4 Algorithm Phase II, achieving arbitrary ϵ accuracy

Given $\widehat{\rho}$ and $\widehat{\Delta}$ from the Phase I of our algorithm. Under the $\Omega(1)$ separation assumptions, we refine the estimation to achieve arbitrary ϵ accuracy in Phase II. In this section, we verify the steps of Algorithm 2, and show the correctness of Theorem 2.2.

4.1 Construct an anchor partition

Imagine that we have a way to group the M words in the vocabulary into a new vocabulary with a *constant* number of superwords, and similarly define marginal vector ρ_A and separation vector Δ_A over the superwords. The new probability matrix (of constant size) corresponds to we sum over the rows/columns of the matrix $\mathbb{B}$ according to the grouping. If we group the words in a way such that Δ_A still has $\Omega(1)$ dictionary separation, then with $N = \Omega(M)$ samples we can estimate the constant-dimensional vector ρ_A and Δ_A to arbitrary accuracy (as $M \gg 1$). Note that accurate estimates of ρ_A and Δ_A defined over the superwords give us crude "global information" about the true ρ and Δ. Now sum the empirical B_N over the rows accordingly, and leave the columns intact, it is easy to recognize that the expected factorization is given by $\rho_A \rho^\top + \Delta_A \Delta^\top$. Therefore, given accurate ρ_A and Δ_A, refining $\widehat{\rho}$ and $\widehat{\Delta}$ is as simple as solving a least square problem.

We formalize this argument and introduce the definition of anchor partition below.

Definition 4.1 (Anchor partition). *Consider a partition of the vocabulary into $(\mathcal{A}, \mathcal{A}^c)$. denote $\rho_{\mathcal{A}} = \sum_{i\in\mathcal{A}} \rho_i$ and $\Delta_{\mathcal{A}} = \sum_{i\in\mathcal{A}} \Delta_i$. We call it an anchor partition if for some constant $C_A = O(1)$,*

$$cond\left(\begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1-\rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix}\right) \le C_A. \tag{25}$$

In this section, we show that if the dictionary separation $\|\Delta\|_1 = \Omega(1)$, the estimator $\widehat{\Delta}$ obtained in Phase I with constant ℓ_1 accuracy contains enough information for us to construct an anchor partition.

First note that with $\Omega(1)$ separation, it is feasible to find a pair of anchor partition. The next lemma state a sufficient condition for constructing an anchor partition.

Lemma 4.2 (Sufficient condition for constructing an anchor partition). *Consider a set of words $\mathcal{I}$, let $\Delta_{\mathcal{I}}$ be the vector of Δ restricted to the entries in $\mathcal{I}$. Suppose that for some constant $C = \Omega(1)$, we have*

$$\|\Delta_{\mathcal{I}}\|_1 \ge C\|\Delta\|_1, \tag{26}$$

and suppose that for some constant $C' \le \frac{1}{3}C$, we can estimate $\Delta_{\mathcal{I}}$ up to precision:

$$\|\widehat{\Delta}_{\mathcal{I}} - \Delta_{\mathcal{I}}\|_1 \le C'\|\Delta_{\mathcal{I}}\|_1. \tag{27}$$

Denote $\widehat{\mathcal{A}} = \{i \in \mathcal{I} : \widehat{\Delta}_i > 0\}$. We have that $(\widehat{\mathcal{A}}, \mathcal{M}\backslash\widehat{\mathcal{A}})$ forms an anchor partition defined in 4.1.

Consider the heaviest bin. If $W_{\log} = \Omega(1)$ and $\|\Delta_{\log}\|_1 = \Omega(1)$, then Lemma 3.3 shows that we can estimate the portion of Δ restricted to the heaviest bin to constant ℓ_1 norm accuracy, and then Lemma 4.2 above shows how to find an anchor partition with set $\mathcal{A}$ as a subset of $\widehat{\mathcal{I}}_{\log}$.

If either $W_{\log} = o(1)$ or $\|\Delta_{\log}\|_1 = o(1)$, there must be at least a constant fraction of marginal probabilities as well as a constant fraction of separation located in the moderate empirical bins ($\widehat{\mathcal{I}}_k$'s). Recall that we can always ignore the lightest bin. If this is the case, in the following we show how to construct an anchor partition using the moderate empirical bins.

Definition 4.3 (Good bin). *Denote the sum of dictionary separation restricted to the "good words"(defined in (12)) $\widehat{\mathcal{L}}_k$ by:*

$$S_k = \sum_{i\in\widehat{\mathcal{L}}_k} |\Delta_i|, \quad \text{for } k = 0, \ldots, \log\log(M). \tag{28}$$

Note that equivalently we can write $S_k = \|\widetilde{\Delta}_k\|_1$.

We call an empirical bin $\widehat{\mathcal{I}}_k$ to be a "good bin" if for some fixed constant $C_1, C_2 = \Omega(1)$ it satisfies the following two conditions:

1. *the marginal probability of the bin $W_k \ge C_1 e^{-k}$.*
2. *the ratio between the separation and the marginal probability of the bin satisfies $\frac{S_k}{2W_k} \ge C_2$.*

The next Lemma shows that Phase I algorithm provides a good estimate of the separation restricted to the good bins with high probability.

Lemma 4.4 (Estimate the separation restricted to the k-th good bin). *Consider the k-th empirical bin $\widehat{\mathcal{I}}_k$ with M_k words. If it is a "good bin", then with high probability, the estimate for the separation vector Δ restricted to the k-th empirical bin $\widehat{\Delta}_{\widehat{\mathcal{I}}_k} = v_k$ given in Lemma 3.6, up to accuracy:*

$$\|\widehat{\Delta}_{\widehat{\mathcal{I}}_k} - \widetilde{\Delta}_k\|_1 \le \frac{1}{\sqrt{d_0}}\|\Delta_{\widehat{\mathcal{I}}_k}\|_1. \tag{29}$$

Let $\mathcal{G} \subseteq \{1, \ldots, \log\log(M)\}$ denote the set of all the "good bins". Next, we show that there are many good bins when the dictionary separation is large.

Lemma 4.5 (Most words fall in "good bins" with high probability)**.** *Assume that* $\|\Delta_{\widehat{\mathcal{I}}_{\log}}\|_1 < \frac{1}{2}\|\Delta\|_1$. *Fix constants* $C_1 = C_2 = \frac{1}{24}\|\Delta\|_1$. *Note that* $C_1 = C_2 = \Omega(1)$ *by well separation Assumption.*

With high probability, we can bound the total marginal probability mass in the "good bins" as:

$$\sum_{k\in\mathcal{G}} W_k \geq \frac{\|\Delta\|_1}{12}. \tag{30}$$

By definition this implies a bound of total separation in the good words in the good bins:

$$\sum_{i\in\widetilde{\mathcal{L}}_k, k\in\mathcal{G}} |\Delta_i| = \sum_{k\in\mathcal{G}} S_k \geq 2C_2 \sum_{k\in\mathcal{G}} W_k \geq \frac{1}{24}(\|\Delta\|_1)^2 = \Omega(1). \tag{31}$$

Lemma 4.5 together with Lemma 4.4 suggest that we can focus on the estimation of separation vector restricted to the "good words" in the "good bins", namely $\widetilde{\Delta}_{\widehat{\mathcal{I}}_k}$ for all $k \in \mathcal{G}$. In particular, set $\mathcal{I} = \cup_{k\in\mathcal{G}}\widehat{\mathcal{L}}_k$ in Lemma 4.2. By Lemma 4.5 we know the separation contained in $\mathcal{I}$ is at least $\sum_{k\in\mathcal{G}} S_k = C\|\Delta\|_1$; moreover by Lemma 4.4, with linear number of samples (large d_0) we can estimate $\Delta_{\mathcal{I}}$ up to constant ℓ_1 accuracy. Therefore we can construct a valid anchor partition $(\mathcal{A}, \mathcal{M}\backslash\mathcal{A})$ by setting:

$$\mathcal{A} = \{\widehat{\Delta}_i > 0 : k \in \mathcal{G}\}.$$

Ideally, we need to restrict to the "good words" and set the anchor partition to be $\{i \in \widetilde{\mathcal{L}}_k, \widehat{\Delta}_i > 0 : k \in \mathcal{G}\}$. However, empirically we do not know which are the "good words" instead of spillover from heavier bins. Luckily, the bound on the total marginal of spillover $\sum_k \overline{W}_k = O(e^{-d_0})$ guarantees that even if we mis-classify all the spillover words, it does not ruin the separation in $\mathcal{A}$ constructed above.

4.2 Estimate the anchor matrix

Now consider grouping the words into two superwords according to the anchor partition we constructed.

We define the 2×2 matrix $D_{\mathcal{A}} = \begin{bmatrix} \rho_{\mathcal{A}}, & \Delta_{\mathcal{A}} \\ 1-\rho_{\mathcal{A}}, & -\Delta_{\mathcal{A}} \end{bmatrix}$ to be the anchor matrix.

To estimate the anchor matrix, we just need to accurately estimate two scalar variables $\rho_{\mathcal{A}}$ and $\Delta_{\mathcal{A}}$. Apply the standard concentration bound, we can argue that with high probability,

$$\| \begin{bmatrix} \sum_{i\in\mathcal{A}, j\in\mathcal{A}}[B_N]_{i,j} & \sum_{i\in\mathcal{A}, j\in\mathcal{A}^c}[B_N]_{i,j} \\ \sum_{i\in\mathcal{A}^c, j\in\mathcal{A}}[B_N]_{i,j} & \sum_{i\in\mathcal{A}^c, j\in\mathcal{A}^c}[B_N]_{i,j} \end{bmatrix} - D_{\mathcal{A}}D_{\mathcal{A}}^{\top}\| < O(\frac{1}{\sqrt{N}})$$

Moreover we have that $|\Delta_{\mathcal{A}}| = \Omega(1)$ since $(\mathcal{A}, \mathcal{A}^c)$ is an anchor partition. Therefore we can estimate $\rho_{\mathcal{A}}$ and $\Delta_{\mathcal{A}}$ to accuracy $\frac{1}{\sqrt{N}}$, and when $N = \Omega(M)$ and M is asymptotically large, we can essentially obtain a close to exact $D_{\mathcal{A}}$.

4.3 Use anchor matrix to estimate dictionary

Now given an anchor partition of the vocabulary $(\mathcal{A}, \mathcal{A}^c)$, and given the exact anchor matrix $D_{\mathcal{A}}$ which has $\Omega(1)$ condition number, refining the estimation of ρ_i and Δ_i for each i is very easy and achieves optimal rate.

Lemma 4.6 (Estimate ρ and Δ with accuracy in ℓ_2 distance)**.** *We have that with probability at least* $1-\delta$*,*

$$\|\widehat{\rho}-\rho\| < \sqrt{\delta/N}, \quad \|\widehat{\Delta}-\Delta\| < \sqrt{\delta/N}.$$

Corollary 4.7 (Estimate ρ and Δ in ℓ_1 distance)**.** *With large probability we can estimate* ρ *and* Δ *in* ℓ_1 *distance:*

$$\|\widehat{\rho}-\rho\|_1 < \sqrt{\frac{M\delta}{N}}, \quad \|\widehat{\Delta}-\Delta\|_1 < \sqrt{\frac{M}{N}}.$$

5 Sample complexity lower bounds for estimation VS testing

5.1 Lower bound for estimating probabilities

We reduce the estimation problem to the community detection for a specific set of model parameters.

Consider the following topic model with equal mixing weights, i.e. $w = w^c = 1/2$. For some constant $C_\Delta = \Omega(1)$, the two word distributions are given by:

$$p = \left[\frac{1+C_\Delta}{M}, \ldots, \frac{1+C_\Delta}{M}, \frac{1-C_\Delta}{M}, \ldots, \frac{1-C_\Delta}{M}\right],$$
$$q = \left[\frac{1-C_\Delta}{M}, \ldots, \frac{1-C_\Delta}{M}, \frac{1+C_\Delta}{M}, \ldots, \frac{1+C_\Delta}{M}\right].$$

The expectation of the sum of samples is given by

$$\mathbb{E}[B_N] = N\frac{1}{2}(pp^\top + qq^\top) = \frac{N}{M^2}\begin{bmatrix} 1+C_\Delta^2 & 1-C_\Delta^2 \\ 1-C_\Delta^2 & 1+C_\Delta^2 \end{bmatrix}.$$

Note that the expected row sum is in the order of $\Omega(\frac{N}{M})$. When N is small, with high probability the entries of the empirical sum B_N only take value either 0 or 1, and B_N approximately corresponds to a SBM ($G(M, a/M, b/M)$) with parameter $a = \frac{N}{M}(1+C_\Delta^2)$ and $b = \frac{N}{M}(1-C_\Delta^2)$.

If the number of sample document is large enough for any algorithm to estimating the dictionary vector p and q up to ℓ_1 accuracy ϵ for a small constant ϵ, it can then be used to achieve partial recovery in the corresponding SBM, namely correctly classify a γ proportion of all the nodes for some constant $\gamma = \frac{\epsilon}{C_\Delta}$.

According to Zhang & Zhou [52], there is a universal constant $C > 0$ such that if $(a-b)^2/(a+b) < c\log(1/\gamma)$, then there is no algorithm that can recover a γ-correct partition in expectation. This suggests that a necessary condition for us to learn the distributions is that

$$\frac{(2(N/M)C_\Delta^2)^2}{2(N/M)} \geq c\log(C_\Delta/\epsilon),$$

namely $(N/M) \geq c\log(C_\Delta/\epsilon)/2C_\Delta^4$. In the well separated regime, this means that the sample complexity is at least linear in the vocabulary size M.

Note that this lower bound is in a sense a worst case constructed with a particular distribution of p and q, and for other choices of p and q it is possible that the sample complexity can be much lower than that $\Omega(M)$.

5.2 Lower bound for testing property of HMMs

In this section, we analyze the information theoretical lower bound to achieve the task of testing whether a sequence of observations is indeed generated by a 2-state HMM. This problem is closely related to the problem of, in our general setup in the main text, testing whether the underlying probability matrix of the observed sample counts is of rank 1 or rank 2. Note that in the context of HMM this would be a stronger lower bound, since we permit an estimator to have more information given the *sequence* of consecutive observations, instead of merely bigram counts.

Theorem 5.1 (Theorem 1.6 restated)**.** *Consider a sequence of N observations from a HMM with two states $\{+,-\}$ and emission distributions p, q supported on M elements. For asymptotically large M, using a sequence of $N = O(M)$ observations, it is information theoretically impossible to distinguish the case that the two emission distributions are well separated, i.e. $||p-q||_1 \geq 1/2$, from the case that both p and q are uniform distribution over $[M]$, namely a degenerate HMM of rank 1.*

In order to derive a lower bound for the sample complexity, it suffices to show that given a sequence of N consecutive observed words, for $N = o(M)$, one can not distinguish whether it is generated by a random instance from a class of 2-state HMMs (Definition 1.4) with well-separated emission distribution p and q, or the sequence is simply N i.i.d. samples from the uniform distribution over $\mathcal{M}$, namely a degenerate HMM with $p = q$.

We shall focus on a class of well-separated HMMs parameterized as below: a symmetric transition matrix $T = \begin{bmatrix} 1-t, & t \\ t, & 1-t \end{bmatrix}$, where we set the transition probability to $t = 1/4$; the initial state distribution is $\pi_p = \pi_q = 1/2$ over the two states s_p and s_q; the corresponding emission distribution p and q are uniform over two disjoint subsets of the vocabulary, $\mathcal{A}$ and $\mathcal{M}\backslash\mathcal{A}$, separately. Moreover, we treat the set $\mathcal{A}$ as a random variable, which can be any of the $\binom{M}{M/2}$ subset of the vocabulary of size $M/2$, which equal probability $1/\binom{M}{M/2}$. Note that there is a one to one mapping between the set $\mathcal{A}$ and an instance in the class of well-separated HMM.

Now consider a random sequence of N words $G_1^N = [g_1, \ldots, g_N] \in \mathcal{M}^N$. If this sequence is generated by an instance of 2-state HMM denoted by $\mathcal{A}$, the joint probability of $(G_1^N, \mathcal{A})$ is given by:

$$\Pr_2(G_1^N, \mathcal{A}) = \Pr_2(G_1^N|\mathcal{A})\Pr_2(\mathcal{A}) = \Pr_2(G_1^N|\mathcal{A})\frac{1}{\binom{M}{M/2}} \tag{32}$$

Moreover, given $\mathcal{A}$, since the support of p and q are disjoint over $\mathcal{A}$ and $\mathcal{M}\backslash\mathcal{A}$ by our assumption, we can perfectly infer about the sequence of hidden states $S_1^N(G_1^N, \mathcal{A}) = [s_1, \ldots, s_N] \in \{s_p, s_q\}^N$ simply by the rule $s_i = s_p$ if $g_i \in \mathcal{A}$ and $s_i = s_q$ otherwise. Thus we have:

$$\Pr_2(G_1^N|\mathcal{A}) = \Pr_2(G_1^N, S_1^N|\mathcal{A}) = \frac{1/2}{M/2}\prod_{i=2}^{M}\frac{(1-t)\mathbf{1}[s_i = s_{i-1}] + t\mathbf{1}[s_i \neq s_{i-1}]}{M/2}. \tag{33}$$

On the other hand, if the sequence G_1^N is simply i.i.d. samples from uniform distribution over $\mathcal{M}$, its probability is given by

$$\Pr_1(G_1^N) = \frac{1}{M^N}. \tag{34}$$

We further define a joint distribution rule $\Pr_1(G_1^N, \mathcal{A})$ such that the marginal probability agrees with $\Pr_1(G_1^N)$. In particular, we define:

$$\Pr_1(G_1^N, \mathcal{A}) = \Pr_1(\mathcal{A}|G_1^N)\Pr_1(G_1^N) \equiv \frac{\Pr_2(G_1^N|\mathcal{A})}{\sum_{\mathcal{B}\in\binom{M}{M/2}}\Pr_2(G_1^N|\mathcal{B})}\Pr_1(G_1^N), \tag{35}$$

where we define the conditional probability $\Pr_1(\mathcal{A}|G_1^N)$ using the properties of the 2-state HMM class.

The main idea of the proof to Theorem 5.1 is to show that for $N = o(M)$, the total variation distance between $\Pr_1$ and $\Pr_2$, is small. It follows immediately from the connection between the error bound of hypothesis testing and total variation distance between two probability rules, that if $TV(\Pr_1(G_1^N), \Pr_2(G_1^N))$ is too small we are not able to test which probability rule the random sequence G_1^N is generated according to.

The detailed proofs are provided in Appendix D.

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. *arXiv preprint arXiv:1503.00609*, 2015.

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

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

[5] Anima Anandkumar, Yi kai Liu, Daniel J. Hsu, Dean P Foster, and Sham M Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25*. 2012.

[6] 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.

[7] 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.

[8] 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.

[9] 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. *arXiv preprint arXiv:1502.03520*, 2015.

[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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] B. Bhattacharya and G. Valiant. Testing closeness with unequal sized samples. In *Neural Information Processing Systems (NIPS) (to appear)*, 2015.

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

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

[18] Yuxin Chen and Emmanuel J Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *arXiv preprint arXiv:1505.05114*, 2015.

[19] 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. *arXiv preprint arXiv:1501.05021*, 2015.

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

[21] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: optimal algorithms via reductions. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2013.

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

[23] 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.

[24] 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.

[25] Peter W Glynn. Upper bounds on poisson tail probabilities. *Operations research letters*, 6(1):9–14, 1987.

[26] 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.

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

[28] 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.

[29] 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.

[30] H. K. Jankowski and J. A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 2009.

[31] 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.

[32] 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.

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

[34] Can M Le and Roman Vershynin. Concentration and regularization of random graphs. *arXiv preprint arXiv:1506.00669*, 2015.

[35] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*. 2014.

[36] 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.

[37] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[38] 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.

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

[40] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

[41] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

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

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

[50] 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.

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

[52] Anderson Y Zhang and Harrison H Zhou. Minimax rates of community detection in stochastic block model.