NEW TECHNIQUES FOR LEARNING AND INFERENCE IN BAYESIAN MODELS

Andrej Risteski

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF COMPUTER SCIENCE

Adviser: Sanjeev Arora

November 2017

ABSTRACT

A common theme in machine learning is succinct modeling of distributions over large domains. Probabilistic graphical models are one of the most expressive frameworks for doing this. The two major tasks involving graphical models are learning and inference. Learning is the task of calculating the best fit graphical model from raw data, while inference is the task of answering probabilistic queries for a known graphical model (for example what is the marginal distribution of one of the variables or what is the distribution of a subset of variables, after conditioning on the values of some other subset of variables). Learning can be thought of as finding a graphical model that explains the raw data, while the inference queries extract the knowledge the graphical model contains.

This thesis introduces new provable techniques for performing both of these tasks, in the context of both latent-variable models – in which a portion of the variables in the graphical model are not observed, as well as fully-observable undirected graphical models (Markov Random Fields). Chapters 2 and 3 will focus on learning latent-variable models, while Chapter 4 will focus on inference in Markov Random Fields.

In Chapter 2, I will contribute the first provable results for analyzing variational Bayes: a family of alternating-minimization style algorithms which is very popular in practice for learning latent-variable models. Despite it's popularity with practitioners, the only theoretical guarantees prior to this work concerned convergence to local minima. We will prove that under reasonable assumptions, in the context of topic models, these algorithms will converge to the global minimum.

Subsequently, in Chapter 3, we will use the method-of-moments along with new techniques in tensor decomposition and constrained matrix factorization to derive algorithms for learning noisy-OR networks – the textbook example of a probabilistic model for causal relationships. Importantly, these techniques were only applicable to linear latent-variable models – which noisy-OR is not.

In Chapter 4, I will contribute a new understanding of a class of variational methods for calculating partition functions in Markov Random Fields. The key technical ingredient is a connection to convex programming hierarchies – a recent area of interest in combinatorial optimization, along with approximations of the entropy of a distribution based on low-order moment information.

Acknowledgements

I thank my advisor, Sanjeev Arora, for his advice, encouragement and mentorship. His view on mathematics and science, passion and energy have surved as ideals towards which to aspire. In particular, his approach to choosing problems on which to work was life-changing for me, and brought about many of the papers included in this thesis. Aside from seemingly having infinite time to do research, he was immensely helpful in making me a better writer and speaker, which were skills I did not have much penchant for or interest in.

Princeton was the perfect atmosphere for collaborations and doing research. I thank all of my collaborators: Pranjal Awasthi, Moses Charikar, Yingyu Liang, Yuanzhi Li, Elad Hazan, Tengyu Ma, and Rong Ge for their willingness to discuss research, and helping me improve my technical writing. I feel like I made not only colleagues, but life friends in the course of our interactions. Yuanzhi Li was also a great officemate, always willing to do research, and always on the lookout for interesting problems.

I would also like to thank all the administrators in the Computer Science department for their impeccable organizational skills (of which I have none.) In particular, Mitra Kelly, Melissa Lawson and Nicki Gotsis saved me from missing more than one deadline in my career.

Finally, I thank my parents, grandparents and relatives which have been supportive of me, despite the fact that I chose a life-path of living on the opposite side of the world from them since the age of 18. In particular, I dedicate this dissertation to both of my late grandfathers, Ilija and Gjorgija, both of whom I miss very much.

Contents

Chapter 1

Introduction to Bayesian modeling

One of the most pervasive modeling paradigms in machine learning is succinctly describing distributions over very large configuration spaces. To list a few examples, consider for instance image segmentation, in which we wish to assign a probability to any configuration of foreground/background values for the pixels in the image – this is a distribution over a configuration space exponential in the number of pixels. Another example is modeling causal relations in data, e.g. diseases and symptoms in a patient – the configuration space here is all possible configurations of presence/absence of symptoms and diseases – again, exponentially sized. In more modern applications, the probabilistic relation between a good "representation" of the data and the data itself is often modeled as a structured, concise probabilistic model like a Restricted Boltzmann Machine (RBM), Deep Boltzmann Machine (DBM), etc.

The desire for succinct models is manifold: often parsimony in the model ensures the parameters capture some *meaningful structure* in the data; frequently the parameters of these models are learned from data, and parsimony helps with the amount of data required; finally, following Occam's razor considerations – maximizing entropy, subject to simple constraints on the distributions (i.e. making minimal assumptions, beyond the constraints we have), often leads to succinct models.

Approaches in machine learning which assign probabilities to data points are typically referred to as *Bayesian* or *generative*, and the focus of this thesis will be on developing new algorithms with provable guarantees for performing the two most common tasks involving Bayesian models: *learning* and *inference*. In this chapter, we will briefly review the basic definitions and tasks involving Bayesian models. Concretely, in Section 1.1 we review the definitions of latent variable, as well as fully observable graphical models. In Section 1.2 we review formally learning and inference of graphical models, as well as common approaches for performing these tasks, both in practice and in theory.

1.1 Latent-variable and fully observed Bayesian models

We will consider two types of Bayesian models in this thesis: latent-variable, directed graphical models and fully observed, undirected graphical models. We proceed to formally define each one of the two.

1.1.1 Latent-variable models

Latent-variable graphical models capture the paradigm of the observable data being "simple", conditioned on some *latent* (unobserved) variables. Mathematically, these models have latent variables $h \in H$ (for some domain H), and observable variables $x \in X$ (for some domain X). The joint distribution of h, x is a function of some *model parameters* θ, and is described as

$$p_\theta(x, h) = p_\theta(h)p_\theta(x|h) \tag{1.1.1}$$

where both terms $p_\theta(h)$ and $p_\theta(x|h)$ are simple functions. Some classical examples that fit in this framework are *mixture models* (e.g. Gaussian mixture models), *topic models*, Bayesian networks (a classical example of which is *noisy-OR* networks and Deep Belief Networks). We will describe in detail topic models and noisy-OR networks, as we will be exhibiting multiple results involving them throughout this thesis.

Topic models are used for modeling *topic structure* in text corpora (Blei and Lafferty, 2009). The model postulates that the expected frequency of words in a document can be written as a mixture of *topic* frequencies – which are themselves distributions over words.

More precisely, the samples x are documents, which are the list of words in the document. (If the vocabulary size is n, and the length of the document is N, we can identify the domain naturally as $[n]^N$.) The latent variables in the document are of two types: the *topic proportions* $\gamma_i, i \in [k]$ of the topics in the document (where k is the number of topics in the corpus), and the *topic* z_j at each position $j \in [N]$ of the document.

The parameters of the model are the distributions of words for each topic $i \in [k], \beta_{i,j}, j \in [n]$, as well a *prior* α for the topic distributions in the document. With this, the procedure for generating a sample x is simple:

- Sample a proportion of topics $\gamma_1, \gamma_2, \ldots, \gamma_k$ according to α.
- For each position l in the document, pick a topic z_l according to a multinomial distribution with parameters $\gamma_1, \ldots, \gamma_k$.
- Conditioned on topic i being picked at that position, pick a word j from a multinomial with parameters $(\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,k})$

The graphical model corresponding to the above procedure is illustrated in Figure 1.1.1, and the sequential nature of picking the topic proportions, topics and words makes it clear why these models are called *directed*.

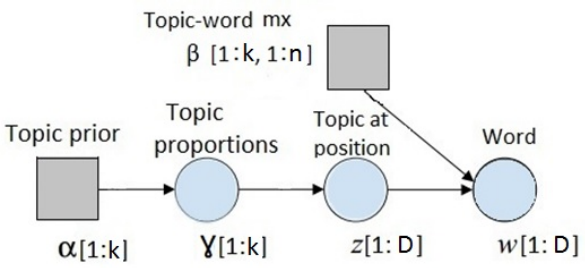


Figure 1.1: Bayesian network corresponding to topic models

Noisy-OR networks are used for modeling causal structure, and are most classically applied to the scenario of diseases and symptoms and this is the nomenclature we will use in this thesis. The model postulates that the symptoms a patient exhibits is a noisy union of the symptoms caused by the diseases a patient has.

More precisely, the samples x are patients, which are 0-1 vectors of dimensions equal to the number of symptoms, indicating whether the patient exhibits that symptom. The latent variable is a 0-1 vector of dimension equal to the number of diseases, indicating whether the patient has that disease.

The parameters of the model are weights $W_{i,j}$, for all symptom/disease pairs i, j, and the prior probabilities ρ_i that a disease i is present. With this, the procedure for generating a sample x is again, simple:

- Sample a disease vector d, s.t. $d_i = 1$ with probability ρ_i, for all diseases i.
- Sample a symptom vector x, s.t. $x_i = 1$ with probability $1 - \exp(\langle W^i, d \rangle)$, where W^i is the vector $W_{i,j}$, for all diseases j.

The name "noisy-or "derives from the fact that a symptom is present if at least one disease which causes it is present, but with some noise "injected" by the above process.

1.1.2 Fully-observed, undirected graphical models

Fully-observable, undirected graphical models (also called *Markov Random Fields*) are used for modeling distributions when the conditional dependency structure is naturally described as a graph. Formally, an undirected graphical model is a distribution of the type

$$p(x) \propto \exp\left(\sum_{c \in \mathrm{cl}(G)} \phi_c(x_c) \right) \tag{1.1.2}$$

where $G = (V, E)$ is a graph, $\mathrm{cl}(G)$ is the set of cliques in G, $\phi_c(\cdot)$ is a *potential* corresponding to the clique c, and x_c is the set of coordinates in x corresponding to the vertices in c. While this most general form might seem somewhat strange, for distributions p which assign positive probability to any configuration, it is equivalent to the so-called *local Markov property*, namely that for any vertex i, x_i is conditionally independent of all other variables, given the

neighborhood of vertex i.

Though the methods we will discuss are fully general, for simplicity we will focus on the classical case of Ising models, which are a *pairwise interaction* undirected graphical model, where the domain of the distribution is the discrete hypercube $\{-1, 1\}^n$, and the potentials are non-zero along the edges of the graph only, and they are simply the scaled products of the values of the vertices. Formally, an *Ising model* has the form

$$p(x) \propto \exp\left(\sum_{(i,j)\in E(G)} J_{i,j} x_i x_j\right) \tag{1.1.3}$$

for some potentials $J_{i,j}$.

1.2 The two main tasks with graphical models

There are two major tasks involving graphical models: learning and inference. Learning is the task of calculating the best fit model from raw data, while inference is the task of answering probabilistic queries for a known model (for example what is the marginal distribution of one of the variables or what is the distribution of a subset of variables, after conditioning on the values of some other subset of variables). Learning can be thought of as finding parameters that best "explain" the raw data, while the inference queries extract the "knowledge" the model contains.

We will survey the main difficulties and approaches of these two tasks, both in terms of provable works, and in terms of heuristics used in practice.

1.2.1 Learning Bayesian models

In context of learning, we will focus on latent-variable models. The problem is as follows:

LEARNING BAYESIAN MODELS, informally: we are given N samples generated according to a latent-variable model (1.1.1), whose parametric form is known (e.g. a topic model or noisy-OR network), but the parameters of which are unknown: we wish to (approximately) recover the values of the parameters. We note this is what is usually called *properly* learning the model; often times it suffices to learn a distribution for the samples which is close to $p_\theta(x)$, but does not necessarily even have the same parametric form: this is called *improper* learning – we will not discuss this in this thesis.

An important choice in the above discussion is whether we are interested in statistical efficiency only (minimizing the number of samples necessary as a function of the target closeness)or computational efficiency as well? Historically, statisticians studied simpler models, where computational efficiency was not an issue – so statistical efficiency received substantially more attention. Let us denote by $\hat{p}$ is the uniform distribution over the samples. In a classical paper by

(Wilks, 1938), it is shown that the *maximum likelihood estimator*

$$\hat{\theta} = \text{argmax}_{\theta} \mathbb{E}_{x \sim \hat{p}} \log p(x) \tag{1.2.1}$$

satisfies two important properties:

- *Consistency (un-biasedness)*: In the limit $N \to \infty$, $\hat{\theta} \to \theta$.
- *Statistical efficiency*: Among all consistent estimators $\hat{\theta}$, it has asymptotically the smallest *mean-squared* error $\mathbb{E}_{\theta}(\hat{\theta} - \theta)^2$.

Maximum likelihood has lingered on in machine learning since, though with the models being substantially more complex, the maximum-likelihood estimator has an important flaw: the optimization problem (1.2.1) is typically NP-hard, if the samples and model parameters are allowed to be chosen in a worst-case fashion. (Note, the preferable hardness result would be one in which the samples "come from the model". These kinds of average-case results are however beyond the current tools of computational complexity.)

The intuitive reason for the hardness is simple: to solve the optimization problem (1.2.1), we have to optimize over the values of the latent variables as well (or more precisely, the *posterior* distributions over the latent variables). This typically renders (1.2.1) non-convex. In practice, such optimization problems are nevertheless frequently solved in a rather natural way: *alternatingly* optimizing the values of either the model parameters and the posterior distributions over the latent variables, while keeping the other fixed. This is the basis of a wide class of algorithms usually referred to as *Variational Bayes*[1], the simplest one of which is Expectation-Maximization (EM).

We will review the basics of EM and variational Bayes in Section 2.1, and subsequently in Section 2.2 give provable results for these algorithms in the case of topic models.

The computational hardness of maximum-likelihood has motivated considering alternate estimators, especially in approaches with theoretical guarantees. In particular, there has been a Renaissance in provable results using the *method of moments*. The method moments is in fact even older than maximum likelihood (Pearson, 1894), but was abandoned early on because it contrast to max-likelihood, it is neither consistent nor statistically efficient. Due to advances in tensor decomposition algorithm, however, there has been a surge of *provably polynomial-time* algorithms for this framework, assuming reasonable structural assumptions on the parameters of the model.

The principle behind the method of moments is to write expressions for the higher-order *moments* of the random variable x as a function of the model parameters, and solve the system of equations for the model parameters, when plugging in the empirical values for the moments. More formally, the method of moments performs the following steps:

[1] Sometimes they are also called *variational inference*, *variational EM*

(1) Express the moments of x as functions of the model parameters:

$$\mathbb{E}\left[x^{\otimes k}\right] = f_\theta(x) \tag{1.2.2}$$

for some function f_θ depending on the model parameters. (Note: $\mathbb{E}\left[x^{\otimes k}\right]$ is an order k tensor.)
(2) Calculate the empirical moments $\mathbb{E}_{x\sim\hat{p}}\left[x^{\otimes k}\right]$, and solve the system of equations (3.1.1) for θ.

The difficulty in this approach is obviously step (2): for complicated models, the function f_θ will be highly non-linear, so solving the system of equations is a non-trivial matter. We will review instances where this *can* be done using very recent machinery on tensor decomposition in Section 3.1 in the setting of mixture models (e.g. mixtures of Gaussians, topic modeling, etc.), and we will show how how this can adapted to a more complicated model: noisy-OR networks.

1.2.2 Inference in Bayesian models

In the context of inference, we will focus on undirected graphical models. The problem of inference can be loosely defined as follows:

INFERENCE IN BAYESIAN MODELS, marginals: we are given as input the parameters of an undirected graphical model, and wish to calculate marginals in the model, e.g. for subsets $S, T \subseteq V(G)$, and configurations $b_S \in X^{|S|}, b_T \in X^{|T|}$, we wish to calculate $p(x_S = b_S | x_T = b_T)$. The simplest version of the problem would just be calculating marginals of the type $p(x_i = b), i \in V(G), b \in X$.

Broadly, there are two approaches to inference. The first is to set up a Markov Chain which has p as it's stationary distribution that mixes rapidly. This allows us to sample from a distribution close to p, and having such samples, estimating the marginals is straightforward. This is an approach that is well-studied in theoretical computer science, especially for graphical models that have connections to statistical physics like Ising models and Potts models (see e.g. (Levin et al., 2009) for a thorough survey of such methods)

The second approach is based on *variational methods*, which proceed as follows. First, it is well-known that calculating marginals of the type we are interested in can be reduced to calculating partition functions. This is a classical result, and we state formally as Lemma 1. Second, we reduce calculating the partition function to an optimization problem over the polytope of probability distributions over the domain of the variables of the graphical model. This is a very old principle, dating as far back as Gibbs. (Ellis, 2012). Since optimization over this polytope is typically intractable however, modern machine learning focuses on strategies to relax this optimization: i.e. by optimizing over some subset of distributions that have a simple form (e.g. product distributions), or relaxing the polytope of distributions and introducing entropy approximations like the Bethe approximation. We review these strategies in great detail

in Section 4.1.

Lemma 1 (Reducing marginals to partition functions, (Jerrum et al., 1986))**.** *There is a polynomial time algorithm, that given an oracle O that takes as input a graph $G = (V, E)$ and potentials $\phi_c, c \in cl(G)$ and outputs the value of the partition function of the corresponding Ising model $Z = \sum_{x \in \{-1,1\}^n} \exp(\sum_{c \in cl(G)} \phi_c(x_c))$, can calculate the values of any marginals $p(x_S = b_S), S \subseteq V(G), b_S \in X^{|S|}$,*

Chapter 2

Provable guarantees for iterative techniques for learning latent-variable models

In this chapter, we will present new results on provable guarantees for iterative techniques for latent-variable models.

First, in Section 2.1 we will review the basic theory behind deriving iterative algorithms like Expectation-Maximization (EM) and variational Bayes. Subsequently, in Section 2.2 we will present new techniques for analyzing such algorithms in the context of topic models: a well-known and widely used latent-variable model for modeling topical structure of text corpora. More concretely, we will first prove that the standard variational Bayes updates can be proven to work under natural structural assumptions (in Sections 2.3 and 2.4). These sections are based on work in (Awasthi and Risteski, 2015).

Subsequently, we will prove that minor changes to the standard updates can lead to similarly cheap algorithms that also use iterative updates, but that work provably under even weaker structural assumptions in Section 2.8. This section is based on work in (Li et al., 2016).

2.1 Background on variational Bayes

In this section, we review the methodological underpinnings of deriving iterative algorithms for learning latent-variable models like variational Bayes. The simplest form of this family of algorithm is Expectation-Maximization (EM). It dates all the way back to (Dempster et al., 1977) and (Sundberg, 1974) in the 70s. As previously mentioned, the algorithm alternatingly keeps either the model parameter or the latent variable posterior estimates fixed, while optimizing for the other. Formally, at iteration t the algorithm maintains estimates $\theta^t, q_x^t(h)$ of the model parameters and the latent variable posteriors for the samples, which are updated as follow:

- E-step: For each sample x, update the current estimate of the posterior distribution over the latent variables:

$$q_x^t(h) = p(h|x, \theta^{t-1}) \tag{2.1.1}$$

- M-step: Update the current estimate of the model parameters:

$$\theta^t = \text{argmax}_\theta \mathbb{E}_{x\sim\hat{p}} E_{q_x^t}[\log p(x, h|\theta)] \tag{2.1.2}$$

The crucial claim about these updates is that after performing an update, the value of the log-likelihood can only increase:

Lemma 2 ((Dempster et al., 1977; Sundberg, 1974)). *In the above setup, it holds that*

$$\mathbb{E}_{x\sim\hat{p}} \log_{\theta_t}(x) \geq \mathbb{E}_{x\sim\hat{p}} \log_{\theta_{t-1}}(x)$$

Two remarks are in order. First, this algorithm assumes that both steps can be efficiently performed. While this is true in some simple models like Gaussian mixture models, even slightly more complicated models like topic models can render these steps intractable. In fact, in topic models, (Sontag and Roy, 2011) showed that calculating the posterior distribution in the E-step, even up to multiplicative polynomial factors, can be #-P hard if the sample x and the model parameters are chosen in a worst-case manner.

Second, in this simple form, there is no "optimization" in the E-step to speak of (it is merely set to some particular value as a function of the current estimate of the model parameters), so it is unclear in which sense the EM algorithm performs "alternating optimization". Towards elucidating this point, we will state a slightly more general result, which will readily lead us to variational Bayes. For any set of distributions $q_x : H \to \mathbb{R}$ indexed by the samples x, we consider the function

$$F(q_x, \theta) = -\mathbb{E}_{x\sim\hat{p}} \text{KL}(q_x || p(h|x)) + \mathbb{E}_{x\sim\hat{p}} \log p(x) \tag{2.1.3}$$

Then, the following claim holds:

Lemma 3 (Alternating minimization view of EM, (Csiszśar and Tusnády, 1984)). *The EM algorithm is equivalent to the following alternating optimization algorithm:*

- *E-step: For each sample x, update the current estimate of the posterior distribution over the latent variables:*

$$q_x^t = argmax_{q_x} F(q_x, \theta^{t-1}) \tag{2.1.4}$$

- *M-step: Update the estimates of the model parameters:*

$$\theta^t = argmax_\theta F(q_x^t, \theta) \tag{2.1.5}$$

The above objective is in machine learning often referred to as the ELBO bound (Blei et al., 2017), but dates back to Gibbs in statistical physics (Ellis, 2012), where it is usually referred to as *variational free energy*. It was written in the above alternating minimization form to elucidate various alternating minimization-like procedures stemming with information-geometric root, which included EM, the Arimoto-Blahut algorithm (Yeung, 2008) (which has made a re-appearence in recent years as a basis of the information bottleneck principle (Tishby et al., 2000)), maximizing returns on investment portfolios etc.

The statistical physics roots naturally lead to variation Bayes methods: even if the optimization problem (2.1.4) is computationally intractable, it may be the case that we can solve an easier minimization problem

$$\text{argmax}_{q_x \in Q} F(q_x, \theta^{t-1}) \tag{2.1.6}$$

for a constrained class of distributions Q. From a theory perspective, even for extremely simple families Q, e.g. product distributions, this problem itself is non-convex, but from a practical perspective, often at least the objective is simple enough as a function of the natural parameters of the distributions, so that the gradient descent updates have a closed form.

Despite this meta-approach being immensely popular in practice, theoretical understanding is very limited. Prior to this thesis, most existing analyses focused on *convergence* results: namely, on showing that the updates eventually reach a fixed point. Exceptions to this are classical results analyzing Lloyd's algorithm for K-means, which is very closely related to the EM algorithm for mixtures of Gaussians (Kumar and Kannan, 2010), (Dasgupta and Schulman, 2000), (Dasgupta and Schulman, 2007). Another line of recent work has focused on a related alternating optimization heuristic in the context of dictionary learning: (Agarwal et al., 2013), (Arora et al., 2015a) prove that with appropriate initialization, alternating minimization can provably recover the ground truth. (Netrapalli et al., 2013) have proven similar results in the context of phase retrieval.

In the following section, we will elucidate more the global convergence properties of variational Bayes: namely, we will exhibit the first characterization of global convergence of variational inference based algorithms in arguably the simplest setting beyond mixture models: topic models (Blei et al., 2003), where the variational class of distributions Q consists of product distributions (i.e. *mean-field*). We show that under *natural assumptions* on the topic-word matrix and the topic priors, along with *natural initialization*, variational inference converges to the parameters of the underlying ground truth model. Going beyond that, in Section 2.8 after, we show that if we step away from the

variational inference updates, but keep the "alternating minimization" paradigm, we can even design algorithms with better convergence properties (in the sense that they work under substantially milder conditions).

2.2 Overview: provable results for variational Bayes in the case of topic models

In this section, we will give an overview of the *mean-field* variational Bayes updates (Subsection 2.2.2) in the case of topic models – the setting in which we will provide provable results, as well as review other works on topic models with provable guarantees (Subsection 2.2.1). Though these works *do not* analyze iterative updates like variational Bayes, we provide this comparison to illustrate the point that the structural assumptions under which we analyze variational Bayes are qualitatively very similar to the assumptions these works make.

2.2.1 Prior work on topic models

The body of work on topic models is vast (Blei and Lafferty, 2009), and we will survey here only the works that have *provable guarantees*. This includes the sequence of works by (Arora et al., 2012a),(Arora et al., 2013a), as well as (Anandkumar et al., 2013), (Ding et al., 2013), (Ding et al., 2014) and (Bansal et al., 2014). (Arora et al., 2012a) and (Arora et al., 2013a) introduced an influential assumption of *anchor words* on the topic-word matrix, which assumes that each topic has a word which appears in that topic, and no other. Subsequent work modified this assumption in various ways: (Anandkumar et al., 2013) assume a certain expansion on the word-topic graph, which says that for any subset S of topics, the number of words in the support of these topics should be at least $|S| + s_{\max}$, where $s_{\max}$ is the maximum support size of any topic. Importantly, neither paper needs any assumption on the topic priors, and can handle (almost) arbitrarily short documents [1].

All the above works are based on the method of moments: an empirical co-occurrence matrix of the words is formed, from which the topic-word matrix is recovered[2]. Our assumptions on the word-topic matrix will be related to the ones in the above works, but importantly, our documents will need to be long, so that the empirical counts of the words are close to their expected counts. Furthermore, we will have to make structural assumptions on the topic priors.

We note that the case where the documents are short seems significantly more difficult. Namely, there are two main elements to analyzing variational inference updates. One is proving the variational approximation to the posterior distribution over topics is not too bad. The second is proving that the updates do actually reach the global optimum. In the

[1] The document length needs to be at least 3.
[2] We expand on this approach in Section 3.1

case of long documents, the posterior distribution is concentrated, so the mean-field variational class can approximate it well – which is decidedly not true when documents are short.

Informally, our assumptions on the topic-word matrix and the topic prior will be as follows:

- The topics will satisfy a weighted expansion property: for any set S of topics of constant size, for any topic i in this set, the probability mass on words which belong to i, and no other topic in S will be large. (Similar to the expansion in (Anandkumar et al., 2013), but only over constant sized subsets.)

- The number of topics per document will be small. Further, the probability of including a given topic in a document is almost independent of any other topics that might be included in the document already. Similar properties are satisfied by the Dirichlet prior, one of the most popular priors in topic modeling. (Originally introduced by (Blei et al., 2003).) The documents will also have a "dominating topic", similarly as in (Bansal et al., 2014).

- For each word j, and a topic i it appears in, there will be a decent proportion of documents that contain topic i and no other topic containing j. These can be viewed as "local anchor documents" for that word-pair topic.

We state informally the main result of this section – see Sections 2.3 and 2.4 for more details.

Theorem (Informal)**.** Under the above mentioned assumptions, popular variants of variational inference for topic models, with suitable initializations, provably recover the ground truth model in polynomial time.

2.2.2 Topic models: variational inference updates and simplifications

2.2.2.1 Review of mean-field variational Bayes for topic models

In this section we briefly review the derivation of the *mean-field* variational Bayes updates for topic modeling, following closely the description in (Blei et al., 2003). Subsequently, we will simplify these updates in the limit of long documents – a slight variant of which we will give provable guarantees for.

As a notational convenience, in the subsequent sections, a superscript of t will denote the value of a variable at the t-th iteration, and a superscript of $*$ denotes the "ground truth" value of the relevant variable.

We will derive the form of both the E-step and the M-step. Proceeding with the E-step first, recall from (2.1.4) that

$$q^t(Z, \gamma) = \text{argmin}_{q \in Q} KL(q(Z, \gamma) || p(Z, \gamma | x, \alpha^t, \beta^t)$$

for a chosen family Q of variational distributions. Denoting by N the length of the document, the *mean-field* variational

family Q we will consider satisfies the constraint that

$$q(\gamma, Z) = q_{\tilde{\gamma}}(\gamma)\Pi_{j=1}^{N} q_{\phi_j}(Z_j)$$

where $q_{\tilde{\gamma}}$ is a Dirichlet distribution with parameters $\tilde{\gamma}$ and $q_{\phi_{j,i}}, j \in [N], i \in [n]$ is a multinomial distribution with parameters $\phi_{j,i}$. In words, this means that distribution q factorizes over the latent variables corresponding to the latent topics Z_j for all positions j, as well as the topic proportions γ. For this variational family, (Blei et al., 2003) show that the *coordinate ascent* equations take a simple form. More precisely,

Lemma 4 ((Blei et al., 2003)). *(1) Keeping $\alpha^t, \beta^t, \phi^t_{j'}, j' \neq j$ fixed, and denoting by w_j the word at position j,*

$$\nabla_{\phi_{j,i}} F(\alpha^t, \beta^t, \gamma^t, \{\phi^t_{j'}, j' \neq j\}) = 0$$

implies

$$\phi_{j,i} \propto \beta^t_{i,w_j} e^{E_q[\log(\gamma_d)|\tilde{\gamma}'_d]}, j \in [N], i \in [n] \tag{2.2.1}$$

$$\tilde{\gamma}_{d,i} = \alpha^t_{d,i} + \sum_{j=1}^{N} \phi^t_{j,i} \tag{2.2.2}$$

(2) For a training set of D documents, indexing as $\gamma^t_{d,i}$ and $\phi^t_{d,j,i}$ the variational parameters corresponding to document d, and by N_d the length of document d, we get

$$\nabla_{\beta_{i,j}} F(\alpha^t, \beta^t, \gamma^t, \{\phi^t_{d,j,i}, \gamma^t_{d,i}, i \in [k], j \in [n]\}) = 0 \tag{2.2.3}$$

implies

$$\beta_{i,j} \propto \sum_{d=1}^{D} \sum_{j'=1}^{N_d} \phi^t_{d,j,i} x_{d,j,j'}, i \in [k], j \in [n]$$

The consequence of the above lemma is an obvious coordinate ascent algorithm, where to perform an E-step, the updates (2.2.1) are iterated until convergence, cycling through the variational parameters $\phi_{j,i}$. The M step is obvious from (2.2.3) for the β parameters. As far as the α Dirichlet parameters are concerned, the gradients with respect to them do not have a closed form expression and (Blei et al., 2003) suggest updating them via a Newton-Rhapson type of procedure.

2.2.2.2 Simplifying the updates in the long document limit

The difficulty with analyzing the above updates is that it isn't clear how to assign an intuitive meaning to the $\tilde{\gamma}$ and ϕ parameters (it's not even clear what one would like them to be ideally at the global optimum.) We will be however

working in the large document limit - which will simplify the updates, and give us a better handle them.

In particular, in the E-step, in the limit $N \to \infty$, the first term in the update equation for $\tilde{\gamma}$ has a vanishing contribution. In this case, we can simplify the E-update as:

$$\phi_{d,j,i} \propto \beta^t_{i,j}\gamma_{d,i}$$

$$\gamma_{d,i} \propto \sum_{j=1}^{N_d} \phi_{d,j,i}$$

Notice, importantly, in the second update we now use variables $\gamma_{d,i}$ instead of $\tilde{\gamma}_{d,i}$, which are normalized such that $\sum_{i=1}^{K} \gamma_{d,i} = 1$. This is an intentional abuse of notation – the γ variables correspond to the max-likelihood topic proportions, given our current estimates $\beta^t_{i,j}$ for the model parameters. The M-step will remain as is - but we will focus on the β only, and ignore the α updates - as the α estimates disappeared from the E updates:

$$\beta^{t+1}_{i,j} \propto \sum_{d=1}^{D} \tilde{f}_{d,j}\gamma^t_{d,i}$$

where $\gamma^t_{d,i}$ is the converged value of $\gamma_{d,i}$ and we denote as $\tilde{f}_{d,j}$ the fractional count of word j in document d (i.e. $\tilde{f}_{d,j} = Count(j)/N_d$, where $Count(j)$ is the number of times word j appears in the document, and N_d is the number of words in the document). In this case, the intuitive meaning of the β^t and γ^t variables is clear: they are estimates of the the model parameters, and the max-likelihood topic proportions, given an estimate of the model parameters, respectively.

An alternative way to view these updates is as follows: first we approximate the posterior distribution $P(Z, \gamma|X, \alpha^t, \beta^t)$ by $P(Z|X, \gamma^*, \alpha^t, \beta^t)$, where γ^* is the max-likelihood value for γ, given our current estimates of α, β, and subsequently setting $P(Z|X, \gamma^*, \alpha^t, \beta^t)$ to be a product distribution. It is intuitively clear that in the large document limit, this approximation should not be much worse than the one in (Blei et al., 2003), as the posterior concentrates around the maximum likelihood value. In fact, our guarantee will apply to finite, but long documents. Finally, we will rewrite the above equations in a slightly more convenient form. Denoting $f_{d,j} = \sum_{i=1}^{K} \gamma_{d,i}\beta^t_{i,j}$, the E-step can be written as: iterate until convergence

$$\gamma_{d,i} \to \gamma_{d,i} \sum_{j=1}^{N} \frac{\tilde{f}_{d,j}}{f_{d,j}}\beta^t_{i,j} \tag{2.2.4}$$

The M-step becomes:

$$\beta^t_{i,j} = \beta^{t-1}_{i,j} \frac{\sum_{d=1}^{D} \frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i}}{\sum_{d=1}^{D} \gamma^t_{d,i}} \tag{2.2.5}$$

where $f^t_{d,j} = \sum_{i=1}^{K} \gamma^t_{d,i}\beta^{t-1}_{i,j}$ and $\gamma^t_{d,i}$ is the converged value of $\gamma_{d,i}$.

2.2.2.3 Alternating KL minimization and thresholded updates

Beyond the long-document simplifications above, we will modify the updates slightly. In fact, similar updates appeared in a paper by (Lee and Seung, 2000) in the context of non-negative matrix factorization. There the authors proved that under these updates $\sum_{d=1}^{D} KL(f_{d,j}^t \| \tilde{f}_{d,j})$ is non-decreasing. More concretely:

Lemma 5 ((Lee and Seung, 2000)). *Consider the objective $\tilde{F}(\gamma^t, \beta^t) = \sum_{d=1}^{D} KL(f_{d,j}^t \| \tilde{f}_{d,j})$. Then,*
(1) Let γ^{t+1} be the result of performing the update (2.2.4) an arbitrary number of times. Then,

$$\tilde{F}(\gamma^{t+1}, \beta^t) \leq \tilde{F}(\gamma^t, \beta^t)$$

Furthermore, if $\gamma^{t+1} = argmin_{\gamma \in \Delta_k} \tilde{F}(\gamma, \beta^t)$, where Δ_K is the K-dimensional simplex

$$\tilde{F}(\gamma^{t+1}, \beta^t) \leq \tilde{F}(\gamma^t, \beta^t)$$

(2) Let β^{t+1} be the result of performing the update (2.2.5). Then,

$$\tilde{F}(\gamma^{t+1}, \beta^{t+1}) \leq \tilde{F}(\gamma^{t+1}, \beta^t)$$

In fact, we remark with respect to the two kinds of updates in (1), iterating the γ updates in the first update is a way to solve this convex minimization problem via a version of gradient descent which makes multiplicative updates, rather than additive updates.

In this section, we will make a modification of the M-step in (2) which is very natural. Intuitively, the update for $\beta_{i,j}^t$ goes over all appearances of the word j and adds the "fractional assignment" of the word j to topic i under our current estimates of the variables β, γ. In the modified version we will only average over those documents d, where $\gamma_{d,i}^t > \gamma_{d,i'}^t, \forall i' \neq i$.

The intuitive reason behind this modification is the following. The variational Bayes updates we are studying work with the KL divergence, which puts more weight on the larger entries. Thus, for the documents in D_i, the estimates for $\gamma_{d,i}^t$ should be better than they might be in the remaining documents. (Of course, since the terms $f_{d,j}^t$ involve all the variables $\gamma_{d,i}^t$, it is not a priori clear that this modification will gain us much, but we will prove that it in fact does.) Formally, we discuss the following three modifications of the updates (we call them suggestively tEM, or thresholded EM) as Algorithms 1 and 2 and 3:

Algorithm 1 KL-tEM

- (E-step) Solve the following convex program for each document d:

$$\min_{\gamma^t_{d,i}} \sum_j \tilde{f}_{d,j} \log(\frac{\tilde{f}_{d,j}}{f^t_{d,j}})$$

s.t.
(1): $\gamma^t_{d,i} \geq 0$, $\sum_i \gamma^t_{d,i} = 1$ and $\gamma^t_{d,i} = 0$ if i does not belong to document d
(M-step) Let D_i to be the set of documents d, s.t. $\gamma^t_{d,i} > \gamma^t_{d,i'}, \forall i' \neq i$.

Set $\beta^{t+1}_{i,j} = \beta^t_{i,j} \frac{\sum_{d \in D_i} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \gamma^t_{d,i}}{\sum_{d \in D_i} \gamma^t_{d,i}}$

Algorithm 2 Iterative tEM

- (E-step) Initialize $\gamma_{d,i}$ uniformly among the topics in the support of document d.
Repeat

$$\gamma_{d,i} = \gamma_{d,i} \sum_{j=1}^{N} \frac{\tilde{f}_{d,j}}{f_{d,j}} \beta^t_{i,j} \tag{2.2.6}$$

until convergence.
(M-step) Same as above.

2.2.3 Initializations

We will consider two different strategies for initialization.

First, we will consider the case where we initialize with the topic-word matrix, and the document priors having the correct support. The analysis of tEM in this case will be the cleanest. While the main focus of this Chapter is variational Bayes, we'll show that this initialization can actually be done for our case efficiently.

Second, we will consider an initialization that is inspired by what the current LDA-c implementation uses. Concretely, we'll assume that the user has some way of finding, for each topic i, a *seed document* in which the proportion of topic i is at least C_l. Then, when initializing, one treats this document as if it were pure: namely one sets $\beta^0_{i,j}$ to be the fractional count of word j in this document. We do not attempt to design an algorithm to find these documents. Importantly, in Section 2.8, we can use this same initialization, but with quantitatively much weaker requirements.

2.3 Case study 1: Sparse topic priors, support initialization

We start with a simple case. As mentioned, all of our results only hold in the long documents regime: we will assume for each document d, the number of sampled words is large enough, so that one can approximate the expected

Algorithm 3 Incomplete tEM

- (E-step) Initialize $\gamma_{d,i}$ with the values gotten in the previous iteration, just perform one step of 2.2.6.
(M-step) Same as before.

frequencies of the words, i.e., one can find values $\gamma^*_{d,i}$, such that $\tilde{f}_{d,j} = (1 \pm \epsilon) \sum_{i=1}^{K} \gamma^*_{d,i} \beta^*_{i,j}$. We'll split the rest of the assumptions into those that apply to the topic-word matrix, and the topic priors. Let's first consider the assumptions on the topic-word matrix. We will impose conditions that ensure the topics don't overlap too much. Namely, we assume:

- *Words are discriminative*: Each word appears in $o(k)$ topics.
- *Almost disjoint supports*: $\forall i, i'$, if the intersection of the supports of i and i' is S, $\sum_{j \in S} \beta^*_{i,j} \leq o(1) \cdot \sum_j \beta^*_{i,j}$.

We also need assumptions on the topic priors. The documents will be sparse, and all topics will be roughly equally likely to appear. There will be virtually no dependence between the topics: conditioning on the size or presence of a certain topic will not influence much the probability of another topic being included. These are analogues of distributions that have been analyzed for dictionary learning (Arora et al., 2015a). Formally:

- *Sparse and gapped documents*: Each of the documents in our samples has at most $s = O(1)$ topics. Furthermore, for each document d, the largest topic $i_0 = \text{argmax}_i \gamma^*_{d,i}$ is such that for any other topic i', $\gamma^*_{d,i'} - \gamma^*_{d,i_0} > \rho$ for some (arbitrarily small) constant ρ.
- *Dominant topic equidistribution*: The probability that topic i is such that $\gamma^*_{d,i} > \gamma^*_{d,i'}, \forall i' \neq i$ is $\Theta(1/k)$.
- *Weak topic correlations and independent topic inclusion*: For all sets T with $o(k)$ topics, it must be the case that: $\mathbf{E}[\gamma^*_{d,i} | \gamma^*_{d,i}$ is dominating$] = (1 \pm o(1))\mathbf{E}[\gamma^*_{d,i} | \gamma^*_{d,i}$ is dominating$, \gamma^*_{d,i'} = 0, i' \in T]$. Furthermore, for any set T of topics, s.t. $|T| \leq s - 1$, $\Pr[\gamma^*_{d,i} > 0 | \gamma^*_{d,i'} \forall i' \in T] = \Theta(\frac{1}{k})$

These assumptions are a less smooth version of properties of the Dirichlet prior. Namely, it's a folklore result that Dirichlet draws are sparse with high probability, for a certain reasonable range of parameters. This was formally proven by (Telgarsky, 2013) - though sparsity there means a small number of large coordinates. It's also well known that Dirichlet essentially cannot enforce any correlation between different topics. In fact, we show analogues of the weak topic correlations property and equidistribution in Section 2.6.

The above assumptions can be viewed as a *local* notion of separability of the model, in the following sense. First, consider a particular document d. For each topic i that participates in that document, consider the words j, which only appear in the support of topic i in the document. In some sense, these words are *local anchor words* for that document: these words appear only in one topic of that document. Because of the "almost disjoint supports" property, there will be a decent mass on these words in each document. Similarly, consider a particular non-zero element $\beta^*_{i,j}$ of the topic-word matrix. Let's call D_l the set of documents where $\beta^*_{i',j} = 0$ for all other topics $i' \neq i$ appearing in that document. These documents are like *local anchor documents* for that word-topic pair: in those documents, the word appears as part of only topic i. It turns out the above properties imply there is a decent number of these for any word-topic pair.

Finally, a technical condition: we will also assume that all nonzero $\gamma^*_{d,i}, \beta^*_{i,j}$ are at least $\frac{1}{poly(n)}$. Intuitively, this means if a topic is present, it needs to be reasonably large, and similarly for words in topics. Such assumptions also appear in the context of dictionary learning (Arora et al., 2015a).

We will prove the following:

Theorem 6 ((Awasthi and Risteski, 2015)). *Given an instance of topic modelling satisfying the properties specified above, where the number of documents is $\Omega(\frac{k \log^2 n}{\epsilon^2})$, if we initialize the supports of the $\beta^t_{i,j}$ and $\gamma^t_{d,i}$ variables correctly, after $O(\log(1/\epsilon') + \log n)$ KL-tEM, iterative-tEM updates or incomplete-tEM updates, we recover the topic-word matrix and topic proportions to multiplicative accuracy $1 + \epsilon'$, for any ϵ' s.t. $1 + \epsilon' \leq \frac{1}{(1-\epsilon)^7}$.*

Theorem 7 ((Awasthi and Risteski, 2015)). *If the number of documents is $\Omega(k^4 \log^2 k)$, there is a polynomial-time procedure which with probability $1 - \Omega(\frac{1}{k})$ correctly identifies the supports of the $\beta^*_{i,j}$ and $\gamma^*_{d,i}$ variables.*

Provable convergence of tEM: The correctness of the tEM updates is proven in 3 steps:

- *Identifying dominating topic*: First, we prove that if $\gamma^t_{d,i}$ is the largest one among all topics in the document, topic i is actually the largest topic.
- *Phase I: Getting constant multiplicative factor estimates*: After initialization, after $O(\log n)$ rounds, we will get to variables $\beta^t_{i,j}$, $\gamma^t_{d,i}$ which are within a constant multiplicative factor from $\beta^*_{i,j}$, $\gamma^*_{d,i}$.
- *Phase II (Alternating minimization - lower and upper bound evolution)*: Once the β and γ estimates are within a constant factor of their true values, we show that the lone words and documents have a *boosting* effect: they cause the multiplicative upper and lower bounds to improve at each round.

The updates we are studying are multiplicative, not additive in nature, and the objective they are optimizing is non-convex, so the standard techniques do not work. The intuition behind our proof in Phase II can be described as follows. Consider one update for one of the variables, say $\beta^t_{i,j}$. We show that $\beta^{t+1}_{i,j} \approx \alpha\beta^*_{i,j} + (1-\alpha)C^t\beta^*_{i,j}$ for some constant C^t at time step t. α is something fairly large (one should think of it as $1 - o(1)$), and comes from the existence of the local anchor documents. A similar equation holds for the γ variables, in which case the "good" term comes from the local anchor words. Furthermore, we show that the error in the $\approx$ decreases over time, as does the value of C^t, so that eventually we can reach $\beta^*_{i,j}$. The analysis bears a resemblance to the *state evolution* and *density evolution* methods in error decoding algorithm analysis - in the sense that we maintain a quantity about the evolving system, and analyze how it evolves under the specified iterations. The quantities we maintain are quite simple - upper and lower multiplicative bounds on our estimates at any round t.

Initialization: Recall the goal of this phase is to recover the supports - i.e. to find out which topics are present in a document, and identify the support of each topic. We will find the topic supports first. This uses an idea inspired

by (Arora et al., 2014) in the setting of dictionary learning. Roughly, we devise a test, which will take as input two documents d, d', and will try to determine if the two documents have a topic in common or not. The test will have no false positives, i.e., will never say YES, if the documents don't have a topic in common, but might say NO even if they do. We then ensure that with high probability, for each topic we find a pair of documents intersecting in that topic, such that the test says YES.

We proceed to making these intuitions formal in the coming sections.

2.3.1 Provable convergence of tEM: Proof of Theorem 6

As a reminder, the outline of the proof will be the following.

- *Identifying dominating topic*: For the modified tEM updates, we need to make sure that the topic with maximal $\gamma^t_{d,i}$ is the dominant.
- *Phase I: Getting constant multiplicative factor estimates*: First, we'll show that after initialization, after $O(\log n)$ number of rounds, we will get to variables $\beta^t_{i,j}$, $\gamma^t_{d,i}$ which are within a constant multiplicative factor from $\beta^*_{i,j}$, $\gamma^*_{d,i}$.
  - *Lower bounds on the β and γ variables*: We'll show that determining the supports of the documents and the topic-word matrix, as well as being able to identify the documents in which topic i is large is enough to ensure that all the $\beta^t_{i,j}$ and $\gamma^t_{i,j}$ variables are lower bounded by $\frac{1}{C^0_\beta}\beta^*_{i,j}$ and $\frac{1}{C^0_\gamma}\gamma^*_{d,i}$ respectively for some constants $C^0_\beta \geq 1, C^0_\gamma \geq 1$.
  - *Improving upper bounds on the $\beta^t_{i,j}$ values*: We show that, if the above two properties are satisfied, we can get a multiplicative upper bound of the $\beta^t_{i,j}$ values, which strictly improves at each step until it reaches a constant. This improvement is very fast: we only need a logarithmic number of steps. After this happens, we show that the γ variables corresponding to these β estimates must be within a constant of the ground truth as well.
- *Phase II (Alternating minimization - lower and upper bound evolution)*: Once the β and γ estimates are within a constant factor of their true values, we show that the lone words and documents have a *boosting* effect: they cause the multiplicative upper and lower bounds to improve at each round.

A word about incorporating the "correct supports" assumption in our algorithms. For the β variables this is obvious: we just set $\beta^t_{i,j} = 0$ if $\beta^*_{i,j} = 0$. For the γ variables it's also fairly straightforward. In KL-tEM we mean simply that in the convex program above, we constrain $\gamma^t_{d,i} = 0$ if $\gamma^*_{d,i} = 0$.

In the iterative version, this just means that before starting the γ iterations, we set the initial value to 0 if $\gamma^*_{d,i} = 0$, and uniform among the rest of the variables. Same for the incomplete version.

In the interest of brevity, whenever we say "the supports are correct", the above is what we will mean.

Recall, we use t to count the iterations for β variables. Put another way, $\gamma_{d,i}^t$ is the value we get for $\gamma_{d,i}$ after the β variables were updated to $\beta_{i,j}^t$. (Which of course, implies, $\beta_{i,j}^{t+1}$ will be the values we get for the β variables after the γ variables are updated to $\gamma_{d,i}^t$.)

The proofs are for each of the variants of tEM are similar. For starters, we show everything for KL-tEM, and then just mention how to modify the arguments to get the results for the other variants in section 2.3.1.6.

2.3.1.1 Determining largest topic

First, we show that the "thresholding" operation works. Namely, we show that if $\gamma_{d,i}^t > \gamma_{d,i'}^t, \forall i \neq i'$, then $\gamma_{d,i}^*$ is the largest topic in the document (there is a unique one by the "slightly gapped documents" property). Furthermore, we can say that $\frac{1}{2}\gamma_{d,i}^* \leq \gamma_{d,i}^t \leq 2\gamma_{d,i}^*$.

Lemma 8. *Fix a document d. Let the supports of the γ and β variables be correct. Then, after a γ iteration, if $\gamma_{d,i}^t > \gamma_{d,i'}^t, i \neq i'$, $\gamma_{d,i}^*$ is the largest topic in the document. Furthermore, $\frac{1}{2}\gamma_{d,i}^* \leq \gamma_{d,i}^t \leq 2\gamma_{d,i}^*$.*

Proof. Since there are a constant number of topics in the document, the largest topic has proportion $\Omega(1)$.

Consider the KL-tEM convex optimization problem. The KKT conditions are easily seen to imply[3]:

$$\sum_{j=1}^{n} \frac{\tilde{f}_{d,j}}{f_{d,j}^t} \beta_{i,j}^t = 1 \tag{2.3.1}$$

For each topic i, since we are considering a constrained optimization problem, it has to be the case that it either satisfies 2.3.1, $\gamma_{d,i}^t = 0$ or $\gamma_{d,i}^t = 1$.

Let's assume first that i satisfies 2.3.1. Then,

$$\gamma_{d,i}^t = \sum_{j=1}^{n} \frac{\tilde{f}_{d,j}}{f_{d,j}^t} \beta_{i,j}^t \gamma_{d,i}^t \leq \sum_{j:\beta_{i,j}^* \neq 0} \tilde{f}_{d,j}$$

Let's call the words j, which only appear in the support of topic i in the document *lone* for that topic, and let's denote that set as L_i.

If L_i are the lone words for topic i, $\sum_{j \notin L_i, \beta_{i,j}^* \neq 0} \tilde{f}_{d,j} = so(1) = o(1)$, so

$$\gamma_{d,i}^t \leq \sum_{j \in L_i} (1+\epsilon) \beta_{i,j}^* \gamma_{d,i}^* + o(1) \leq (1+\epsilon)\gamma_{d,i}^* + o(1) \leq \gamma_{d,i}^* + o(1)$$

On the other hand, $\gamma_{d,i}^t \geq \sum_{j \in L_i} \beta_{i,j}^* \gamma_{d,i}^* \geq (1-\epsilon)(1-o(1))\gamma_{d,i}^* \geq (1-o(1))\gamma_{d,i}^*$, so $\gamma_{d,i}^t \geq \gamma_{d,i}^* - o(1)$.

[3] One gets these trivially, turning the constraint that $\sum_{i=1}^{K} \gamma_{d,i}^t = 1$ into a Lagrange multiplier

Since there is a constant gap of ρ between the largest topic and the next largest one, the maximum $\gamma^t_{d,i}$ is indeed the largest topic in the document. Furthermore, since $(1 - o(1))\gamma^*_{d,i} \leq \gamma^t_{d,i} \leq (1 + o(1))\gamma^*_{d,i}$, clearly $\frac{1}{2}\gamma^*_{d,i} \leq \gamma^t_{d,i} \leq 2\gamma^*_{d,i}$ follows as well.

On the other hand, we claim no topic which is in the support of a document d can actually have $\gamma^t_{d,i} = 0$.

If this happens, it's easy to see that $\sum_{j=1}^{n} \tilde{f}_{d,j} \log(\frac{\tilde{f}_{d,j}}{f^t_{d,j}}) = \infty$: one only needs to look at a summand corresponding to a lone word j for topic i. Just by virtue of the way lone words are defined, $\gamma^t_{d,i} = 0$ would imply $f^t_{d,j} = 0$. It's clear that one can get a finite value for $\sum_j \tilde{f}_{d,j} \log(\frac{\tilde{f}_{d,j}}{f^t_{d,j}})$ on the other hand, by just setting $\gamma^t_{d,i} = \gamma^*_{d,i}$, so $\gamma^t_{d,i} = 0$ cannot happen at an optimum.

□

2.3.1.2 Lower bounds on the $\gamma^t_{d,i}$ and $\beta^t_{i,j}$ variables

Next, we show that subject to the thresholding being correct, at any point in time t, all the estimates $\gamma^t_{d,i}$ and $\beta^t_{i,j}$ are appropriately lower bounded.

The proof is similar for both the β and γ variables, and both for the KL-tEM and iterative tEM updates, but as mentioned before, we focus on the KL-tEM first.

Lemma 9. *Fix a particular document d. Suppose that the supports of the γ and β variables are correct. Then, $\gamma^t_{d,i} \geq (1 - o(1))\gamma^*_{d,i}$.*

Proof. Multiplying both sides of 2.3.1 by $\gamma^t_{d,i}$, we get

$$\gamma^t_{d,i} = \sum_{j=1}^{n} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \beta^t_{i,j} \gamma^t_{d,i}$$

As above, let's split the above sum in two parts: lone words, and non-lone. Then clearly,

$$\gamma^t_{d,i} \geq \sum_{j \in L_i} (1 - \epsilon) \beta^*_{i,j} \gamma^*_{d,i}$$

For notational convenience, let's denote $\tilde{\alpha} = \sum_{j \in L_i} \beta^*_{i,j}$. Let's estimate $\tilde{\alpha}$. By the assumption on the size of the intersection of topics,

$$\sum_{j \notin L_i} \beta^*_{i,j} \leq so(1) = o(1)$$

.

Hence, $\tilde{\alpha} \geq (1 - \epsilon)(1 - o(1)) = 1 - o(1)$. So, the claim of the lemma holds.

□

The lower bound on the $\beta^t_{i,j}$ values proceeds similarly, but here we will crucially make use of the fact that for the large topics, we have both upper and lower bounds on the $\gamma^t_{d,i}$ values.

Lemma 10. *Suppose that the supports of the γ and β variables are correct. Additionally, if i is a large topic in d, let $\frac{1}{2}\gamma^*_{d,i} \le \gamma^t_{d,i} \le 2\gamma^*_{d,i}$. Then, $\beta^{t+1}_{i,j} \ge \frac{1}{2}(1 - o(1))\beta^*_{i,j}$.*

Proof. Let's call *lone* the documents where $\beta^*_{i',j} = 0$ for all other topics $i' \neq i$ appearing in that document for the topic-word pair (i, j). Let D_l be the set of lone documents. Then, certainly it's true that

$$\beta^{t+1}_{i,j} \ge \beta^t_{i,j} \frac{\sum_{d \in D_l} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \gamma^t_{d,i}}{\sum_{d=1}^{D} \gamma^t_{d,i}}$$

However, for a lone document, $f^t_{d,j} = \gamma^t_{d,i} \cdot \beta^t_{i,j}$ (it's easy to check all the other terms in the summation for $f^t_{d,j}$ vanish, because either $\gamma^t_{d,i'} = 0$ or $\beta^t_{i',j} = 0$). Hence,

$$\beta^{t+1}_{i,j} \ge \frac{\sum_{d \in D_l} (1 - \epsilon) \frac{\gamma^*_{d,i}\beta^*_{i,j}}{\gamma^t_{d,i}\beta^t_{i,j}} \beta^t_{i,j}\gamma^t_{d,i}}{\sum_{d=1}^{D} \gamma^t_{d,i}} = (1 - \epsilon)\beta^*_{i,j} \frac{\sum_{d \in D_l} \gamma^*_{d,i}}{\sum_{d=1}^{D} \gamma^t_{d,i}}$$

However, since the update is happening only over documents where topic i is large, $\gamma^t_{d,i} \le 2\gamma^*_{d,i}$. So, we can conclude

$$\beta^{t+1}_{i,j} \ge (1 - \epsilon)\beta^*_{i,j} \frac{1}{2} \frac{\sum_{d \in D_l} \gamma^*_{d,i}}{\sum_{d=1}^{D} \gamma^*_{d,i}}$$

Let's call $\alpha = \frac{\sum_{d \in D_l} \gamma^*_{d,i}}{\sum_{d=1}^{D} \gamma^*_{d,i}}$, and let's analyze it's value.

By Lemma 55 and Lemma 54,

$$\sum_{d \in D_l} \gamma^*_{d,i} \ge (1 - \epsilon)|D_l|\mathbf{E}[\gamma^*_{d,i} | \gamma^*_{d,i} \text{ is dominating}, \gamma^*_{d,i'} = 0, \forall i' \neq i \text{ s.t. } j \text{ appears in topic } i']$$

$$\sum_{d=1}^{D} \gamma^*_{d,i} \le (1 + \epsilon)|D|\mathbf{E}[\gamma^*_{d,i} | \gamma^*_{d,i} \text{ is dominating}]$$

By the weak topic correlations assumption, then, $\frac{\sum_{d \in D_l} \gamma^*_{d,i}}{\sum_{d=1}^{D} \gamma^*_{d,i}} \ge (1 - o(1)) \frac{|D_l|}{|D|}$.

Furthermore, by the independent topic inclusion property, each of the $o(K)$ topics other than i that word j belongs to appears in a document with probability $\Theta(1/K)$, so the probability that a document which contains topic i contains one of them is $o(1)$, i.e. $\frac{|D_l|}{|D|}$. By Lemma 56, furthermore, $\frac{|D_l|}{|D|} \ge 1 - o(1)$ when $\epsilon = o(1)$. Hence, $\alpha \ge 1 - o(1)$.

Altogether, we get that $\beta^{t+1}_{i,j} \ge \frac{1}{2}(1 - o(1))\beta^*_{i,j}$ as claimed.

□

2.3.1.3 Upper bound on the $\beta^t_{i,j}$ values

Having established a lower bound on the $\beta^t_{i,j}$ variables throughout all iterations, together with the lower bounds on the $\gamma^t_{d,i}$ variables and the good estimates for the large topics, we will be able to prove the upper bound of the multiplicative error of $\beta^t_{i,j}$ keeps improving, until $\beta^t_{i,j} \leq C_\beta \beta^*_{i,j}$, for some constant C_β.

Lemma 11. *Let the β variables have the correct support, and $\beta^t_{i,j} \geq \frac{1}{C_m}\beta^*_{i,j}$, $\gamma^t_{d,i} \geq \frac{1}{C_m}\gamma^*_{d,i}$ whenever $\beta^*_{i,j} \neq 0$, $\gamma^*_{d,i} \neq 0$. Let $\beta^t_{i,j} = C^t_\beta \beta^*_{i,j}$, where $C^t_\beta \geq 4C_m$, and C_m is a constant. Then, in the next iteration, $\beta^{t+1}_{i,j} \leq C^{t+1}_\beta \beta^*_{i,j}$, where $C^{t+1}_\beta \leq \frac{C^t_\beta}{2}$.*

Proof. Without loss of generality, let's assume $C_m \geq 2$. (Since certainly, if the statement of the lemma holds with a smaller constant, it holds with $C_m = 2$.)

We proceed similarly as in the prior analyses. We will split the sum into the portion corresponding to the lone and non-lone documents.

Let's analyze the terms $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i}$ corresponding to the non-lone documents.

Now, $f^t_{d,j} \geq \frac{1}{C_m^2} f^*_{d,j}$, so $\frac{\tilde{f}_{d,j}}{f^t_{d,j}} \leq (1+\epsilon)C_m^2$. Also, $\gamma^t_{d,i} \leq 2\gamma^*_{d,i}$, since topic i is the dominant in document d. Since $C_m \geq 2$, $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i} \leq (1+\epsilon)C_m^3 \gamma^*_{d,i}$.

Also, note that $\sum_{d=1}^D \gamma^t_{d,i} \geq \frac{1}{C_m}\sum_{d=1}^D \gamma^*_{d,i}$, again, since i is the dominant topic.

As usual, let's denote the set of lone documents D_l:

$$\beta^{t+1}_{i,j} \leq (1+\epsilon)C_m \frac{\sum_{d\in D_l}\beta^*_{i,j}\gamma^*_{d,i} + \sum_{d\in D\setminus D_l} C_m^3 \gamma^*_{d,i}\beta^t_{i,j}}{\sum_{d=1}^D \gamma^*_{d,i}}$$

As in the prior proofs, let's denote by $\alpha := \frac{\sum_{d\in D_l}\gamma^*_{d,i}}{\sum_{d=1}^D \gamma^*_{d,i}}$.

As in Lemma 10, $\alpha \geq 1 - o(1)$, so $\beta^{t+1}_{i,j} \leq (1+\epsilon)C_m(\alpha\beta^*_{i,j} + (1-\alpha)C_m^3\beta^t_{i,j})$, which in turn implies that $\frac{\beta^{t+1}_{i,j}}{\beta^*_{i,j}} \leq (1+\epsilon)C_m(\alpha + (1-\alpha)C_m^3 C^t_\beta)$. In order to ensure that $\frac{\beta^{t+1}_{i,j}}{\beta^*_{i,j}} < \frac{C^t_\beta}{2}$, it would be sufficient to prove that

$$(1+\epsilon)C_m(\alpha + (1-\alpha)(C_m^3 C^t_\beta) < \frac{C^t_\beta}{2}$$

which is equivalent to $\alpha > \dfrac{C_m^3 C^t_\beta - \frac{C^t_\beta}{2(1+\epsilon)C_m}}{C_m^3 C^t_\beta - 1}$.

Let's look at the right hand side. As, by assumption, $C^t_\beta \geq 4C_m$, it follows that

$$\frac{C_m^3 C^t_\beta - \frac{C^t_\beta}{2(1+\epsilon)C_m}}{C_m^3 C^t_\beta - 1} \leq \frac{C_m^3 C^t_\beta - \frac{C^t_\beta}{2(1+\epsilon)C_m}}{C_m^3 C^t_\beta - \frac{C^t_\beta}{4C_m}}$$

Hence, the right hand side is upper bounded by

$$\frac{C_m^3 - \frac{1}{2(1+\epsilon)C_m}}{C_m^3 - \frac{1}{4C_m}} = 1 - \frac{\frac{\frac{2}{1+\epsilon}-1}{4C_m}}{C_m^3 - \frac{1}{4C_m}}$$

But, since C_m is bounded by a constant, and $\alpha = 1 - o(1)$, the claim follows.

□

2.3.1.4 Upper bounds on the γ values

Finally, we show that if we ever reach a point where the β values are both upper and lower bounded by a constant, the γ values one gets after the γ step are appropriately upper bounded by a constant. More precisely:

Lemma 12. *Fix a particular document d. Let's assume the supports for the β and γ variables are correct. Furthermore, let $\frac{1}{C_m} \leq \frac{\beta^t_{i,j}}{\beta^*_{i,j}} \leq C_m$ for some constant C_m. Then, $\gamma^t_{d,i} \leq (1 + o(1))\gamma^*_{d,i}$.*

Proof. As in the proof of Lemma 9, let's look at the KKT conditions for $\gamma^t_{d,i}$ into a part corresponding to lone words L_i and non-lone words. Multiplying 2.3.1 by $\gamma^t_{d,i}$ as before,

$$\gamma^t_{d,i} = \sum_{j \in L_i} \tilde{f}_{d,j} + \gamma^t_{d,i} \sum_{j \notin L_i} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \beta^t_{i,j}$$

Again, let $\tilde{\alpha} = \sum_{j \in L_i} \beta^*_{i,j}$.

By Lemma 9, certainly $\gamma^t_{d,i} \geq \frac{1}{C_m}\gamma^*_{d,i}$. Hence, $\frac{\tilde{f}_{d,j}}{f^t_{d,j}} \leq (1+\epsilon)C_m^2$. So we have, $\gamma^t_{d,i} \leq (1+\epsilon)(\tilde{\alpha}\gamma^*_{d,i} + C_m^3(1-\tilde{\alpha})\gamma^t_{d,i})$. In other words, this implies $\gamma^t_{d,i} \leq \frac{(1+\epsilon)\tilde{\alpha}}{1-(1+\epsilon)C_m^3(1-\tilde{\alpha})}\gamma^*_{d,i}$. Since $\tilde{\alpha} = 1 - o(1)$, it's easy to check that $\frac{\tilde{\alpha}}{1-C_m^3(1-\tilde{\alpha})} \leq 1 + o(1)$, which is enough for what we need.

□

So, as a corollary, we finally get:

Corollary 13. *For some $t_0 = O(\log(\frac{1}{\beta^*_{\min}})) = O(\log n)$, it will be the case that for all $t \geq t_0$, $\frac{1}{C^0_\beta} \leq \frac{\beta^*_{i,j}}{\beta^t_{i,j}} \leq C^0_\beta$ for some constant C^0_β and $\frac{1}{C^0_\gamma} \leq \frac{\gamma^*_{d,i}}{\gamma^t_{d,i}} \leq C^0_\gamma$ for some constant C^0_γ.*

This concludes Phase I of the analysis.

2.3.1.5 Phase II: Alternating minimization - upper and lower bound evolution

Taking Corollary 13 into consideration, we finally show that, if the β and γ values are correct up to a constant multiplicative factor, and we have the correct support, we can improve the multiplicative error in each iteration, thus achieving convergence to the correct values.

This portion bears resemblance to techniques like *state evolution* and *density evolution* in the literature for iterative methods for decoding error correcting codes. In those techniques, one keeps track of a certain quantity of the system that's evolving in each iteration. In density evolution, this is the probability density function of the messages that are being passed, in state evolution, it is a certain average and variance of the variables we are estimating.

In our case, we keep track of the "multiplicative accuracy" of our estimates $\gamma^t_{d,i}, \beta^t_{i,j}$. In particular, we will keep track of quantities C^t_γ and C^t_β, such that at iteration t, $\frac{1}{C^t_\beta} \leq \frac{\beta_{i,j}*}{\beta^t_{i,j}} \leq C^t_\beta$ and $\frac{1}{C^t_\gamma} \leq \frac{\gamma_{d,i}*}{\gamma^t_{d,i}} \leq C^t_\gamma$ after the corresponding γ iteration.

We will show that improvement in the quantities C^t_β causes a large enough improvement in the C^t_γ updates, so that after an alternating step of β and γ updates, $C^{t+1}_\beta \leq (C^t_\beta)^{1/2}$.

First, we show that when the β variables are estimated up to a constant multiplicative factor, the constant for the γ values after they've been iterated to convergence is slightly better than the constant for the β values. More precisely:

Lemma 14. *Let's assume that our current iterates $\beta^t_{i,j}$ satisfy $\frac{1}{C^t_\beta} \leq \frac{\beta_{i,j}*}{\beta^t_{i,j}} \leq C^t_\beta$ for $C^t_\beta \geq \frac{1}{(1-\epsilon)^7}$. Then, after iterating the γ updates to convergence, we will get values $\gamma^t_{d,i}$ that satisfy $(C^t_\beta)^{1/3} \leq \frac{\gamma_{d,i}*}{\gamma^t_{d,i}} \leq (C^t_\beta)^{1/3}$.*

Proof. As usual, we will split the KKT conditions for $\gamma^{t+1,t'}_{d,i}$ into two parts: one for the lone, and one for the non-lone words. Let's call the set of lone words L_i, as previously. Then. we have

$$\gamma^t_{d,i} = \sum_{j \in L_i} \tilde{f}_{d,j} + \gamma^t_{d,i} \sum_{j \notin L_i} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \beta^t_{i,j}$$

Again, let $\tilde{\alpha} = \sum_{j \in L_i} \beta^*_{i,j} = o(1)$, as we proved before.

Let's denote as $C^t_\gamma = \max_i(\max(\frac{\gamma^*_{d,i}}{\gamma^t_{d,i}}, \frac{\gamma^t_{d,i}}{\gamma^*_{d,i}}))$.

We claim that it has to hold that $C^t_\gamma \leq (C^t_\beta)^{1/3}$. Assume the contrary, and let $i_0 = \text{argmax}_i(\max(\frac{\gamma^*_{d,i}}{\gamma^t_{d,i}}, \frac{\gamma^t_{d,i}}{\gamma^*_{d,i}}))$.

Let's first assume that $\frac{\gamma^t_{d,i_0}}{\gamma^*_{d,i_0}} = C^t_\gamma$.

By the definition of C^t_γ,

$$\gamma^t_{d,i_0} = \sum_{j \in L_{i_0}} \tilde{f}_{d,j} + \gamma^t_{d,i_0} \sum_{j \notin L_{i_0}} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \beta^t_{i_0,j} \leq (1+\epsilon)(\tilde{\alpha}\gamma^*_{d,i_0} + (1-\tilde{\alpha})(C^t_\beta)^2(C^t_\gamma)^2\gamma^*_{d,i_0})$$

We claim that

$$(1+\epsilon)(\tilde{\alpha} + (1-\tilde{\alpha})(C^t_\beta)^2(C^t_\gamma)^2) \leq (C^t_\gamma)^{1/3} \tag{2.3.2}$$

which will be a contradiction to the definition of C^t_γ.

After a little rewriting, 2.3.2 translates to $\tilde{\alpha} \geq 1 - \frac{\frac{(C^t_\gamma)^{1/3}}{1+\epsilon} - 1}{(C^t_\beta C^t_\gamma)^2 - 1}$. By our assumption on C^t_γ, $C^t_\beta \leq C^3_\gamma$, so the right hand side above is upper bounded by $1 - \frac{\frac{(C^t_\gamma)^{1/3}}{1+\epsilon} - 1}{(C^t_\gamma)^8 - 1}$.

But, Lemma 12 implies that certainly $C_\gamma^t \le C_\gamma^0$, where C_γ^0 is some absolute constant. The function

$$f(c) = \frac{\frac{c^{1/3}}{1+\epsilon} - 1}{c^8 - 1}$$

can be easily seen to be monotonically decreasing on the interval of interest, and hence is lower bounded by $\frac{\frac{(C_\gamma^0)^{1/3}}{1+\epsilon}-1}{(C_\gamma^0)^8-1}$, which is in terms some absolute constant smaller than one. Since $\tilde{\alpha} = 1 - o(1)$. the claim we want is clearly true.

The case where $\frac{\gamma^*_{d,i_0}}{\gamma^t_{d,i_0}} = C_\gamma^t$ is similar. In this case,

$$\gamma^t_{d,i_0} = \sum_{j \in L_{i_0}} \tilde{f}_{d,j} + \gamma^t_{d,i_0} \sum_{j \notin L_{i_0}} \frac{\tilde{f}d,j}{f^t_{d,j}} \beta^t_{i_0,j} \ge (1-\epsilon)(\tilde{\alpha}\gamma^*_{d,i_0} + (1-\tilde{\alpha})\frac{1}{(C_\beta^t)^2 (C_\gamma^t)^2}\gamma^*_{d,i_0})$$

We then claim that

$$(1-\epsilon)(\tilde{\alpha} + (1-\tilde{\alpha})\frac{1}{(C_\beta^t)^2(C_\gamma^t)^2}) \ge \frac{1}{(C_\gamma^t)^{1/3}} \tag{2.3.3}$$

Again, 2.3.3 rewrites to:

$$\tilde{\alpha} \ge \frac{\frac{1}{(1-\epsilon)(C_\gamma^t)^{1/3}} - \frac{1}{(C_\beta^t)^2(C_\gamma^t)^2}}{1 - \frac{1}{(C_\beta^t)^2(C_\gamma^t)^2}} = 1 - \frac{1 - \frac{1}{(1-\epsilon)(C_\gamma^t)^{1/3}}}{1 - \frac{1}{(C_\beta^t C_\gamma^t)^2}}$$

Again, the right hand side above is upper bounded by $1 - \frac{1-\frac{1}{(1-\epsilon)(C_\gamma^t)^{1/3}}}{1-\frac{1}{(C_\gamma^t)^8}}$. But $C_\gamma \in [1, C_\gamma^0]$, and the function $\frac{1-\frac{1}{(1-\epsilon)c^{1/3}}}{1-\frac{1}{c^8}}$ is monotonically increasing, so lower bounded by

$$\frac{1 - \frac{1}{(1-\epsilon)(\frac{1}{(1-\epsilon)^7})^{1/3}}}{1 - \frac{1}{(\frac{1}{(1-\epsilon)^7})^8}} = \frac{1-(1-\epsilon)^{4/3}}{1-(1-\epsilon)^{56}} \ge \frac{1}{42}$$

Hence, $1 - \frac{1-\frac{1}{(1-\epsilon)(C_\gamma^t)^{1/3}}}{1-\frac{1}{(C_\gamma^t)^{32}}}$ is upper bounded by $\frac{41}{42}$. Again, our bound on $\tilde{\alpha}$ gives us what we want.

□

Lemma 15. *Let's assume that our current iterates $\beta^t_{i,j}$ satisfy $\frac{1}{C_\beta^t} \le \frac{\beta_{i,j}*}{\beta^t_{i,j}} \le C_\beta^t$, $C_\beta^t \ge \frac{1}{(1-\epsilon)^7}$, and after the corresponding γ update, we get $\frac{1}{C_\gamma^t} \le \frac{\gamma_{d,i}*}{\gamma^t_{d,i}} \le C_\gamma^t$, where $C_\beta^t \ge (C_\gamma^t)^3$. Then, after one β step, we will get new values $\beta^{t+1}_{i,j}$ that satisfy $\frac{1}{C_\beta^{t+1}} \le \frac{\beta_{i,j}*}{\beta^{t+1}_{i,j}} \le C_\beta^{t+1}$ where $C_\beta^{t+1} = (C_\beta^t)^{1/2}$.*

Proof. The proof proceeds in complete analogy with Lemmas 10 and 11.

Again, let's tackle the lower and upper bound separately. The upper bound condition is:

$$\alpha > \frac{(C_\beta^t C_\gamma^t)^2 - \frac{(C_\beta^t)^{1/2}}{(1+\epsilon)C_\gamma^t}}{(C_\gamma^t C_\beta^t)^2 - 1}$$

Using $C_\beta^t \geq (C_\gamma^t)^3$, we can upper bound the expression on the right by $1 - \frac{\frac{(C_\beta^t)^{1/6}}{1+\epsilon} - 1}{(C_\beta^t)^{8/3} - 1}$. The function $f(c) = \frac{\frac{x^{1/6}}{1+\epsilon} - 1}{x^{8/3} - 1}$ is monotonically decreasing on the interval $[1, C_\beta^0]$ of interest, so because $\alpha = 1 - o(1)$, we get what we want.

Similarly, for the lower bound, we want that

$$\alpha > \frac{\frac{C_\gamma^t}{(C_\beta^t)^{1/2}(1-\epsilon)} - \frac{1}{(C_\gamma^t C_\beta^t)^2}}{1 - \frac{1}{(C_\beta^t C_\gamma^t)^2}}$$

Yet again, using $C_\beta^t \geq (C_\gamma^t)^3$, we get that the right hand side is upper bounded by

$$1 - \frac{1 - \frac{1}{(1-\epsilon)C_\beta^{1/6}}}{1 - \frac{1}{C_\beta^3}}$$

However, the function $f(c) = \frac{1 - \frac{1}{(1-\epsilon)c^{1/6}}}{1 - \frac{1}{c^{8/3}}}$ is monotonically increasing on the interval $[1, C_\beta^0]$, so lower bounded by $\frac{1 - \frac{1}{(1-\epsilon)(\frac{1}{(1-\epsilon)^7})^{1/6}}}{1 - \frac{1}{(\frac{1}{(1-\epsilon)^7})^{8/3}}} = \frac{1-(1-\epsilon)^{1/6}}{1-(1-\epsilon)^{21}} \geq \frac{1}{126}$. Hence, $1 - \frac{1 - \frac{1}{(1-\epsilon)C_\beta^{1/6}}}{1 - \frac{1}{C_\beta^3}}$ is upper bounded by $\frac{125}{126}$, so using the fact that $\alpha = 1 - o(1)$, we get what we want.

□

Putting lemmas 14 and 15 together, we get:

Lemma 16. *Suppose it holds that* $\frac{1}{C^t} \leq \frac{\beta_{i,j}^*}{\beta_{i,j}^t} \leq C^t$, $C^t \geq \frac{1}{(1-\epsilon)^7}$. *Then, after one KL minimization step with respect to the* γ *variables and one* β *iteration, we get new values* $\beta_{i,j}^{t+1}$ *that satisfy* $\frac{1}{C^{t+1}} \leq \frac{\beta_{i,j}^*}{\beta_{i,j}^{t+1}} \leq C^{t+1}$, *where* $C^{t+1} = \sqrt{C^t}$

Proof. By Lemma 14, after the γ iterations, we get $\gamma_{d,i}^t$ values that satisfy the condition $\frac{1}{(C')^t} \leq \frac{\gamma_{d,i}^*}{\gamma_{d,i}^t} \leq (C')^t$, where $(C')^t = (C^t)^{1/3}$.

Then, by Lemma 15, after the γ iteration, we will get $\frac{1}{C^{t+1}} \leq \frac{\beta_{i,j}^*}{\beta_{i,j}^{t+1}} \leq C^{t+1}$, such that $C^{t+1} = (C^t)^{1/2}$, which is what we need.

□

Hence, as a corollary, we get immediately:

Corollary 17. *Lemma 16 above implies that Phase III requires* $O(\log(\frac{1}{\log(1+\epsilon')})) = O(\log(\frac{1}{\epsilon'}))$ *iterations to estimate each of the topic-word matrix and document proportion entries to within a multiplicative factor of* $1 + \epsilon'$.

This finished the proof of Theorem 6 for the KL-tEM version of the updates. In the next section, we will remark on why the proofs are almost identical in the iterative and incomplete tEM version of the updates.

2.3.1.6 Iterative tEM updates, incomplete tEM updates

We show how to modify the proofs to show that the iterative tEM and incomplete tEM updates work as well. We'll just sketch the arguments as they are almost identical as above.

In those updates, when we are performing a γ update, we initialize with $\gamma^t_{d,i} = 0$ whenever topic i does not belong to document d, and $\gamma^t_{d,i}$ uniform among all the other topics.

Then, the way to modify Lemmas 9, 12, 14 is simple. Instead of arguing by contradiction about what happens at the KKT conditions, one will assume that at iteration t' (t' to indicate these are the separate iterations for the γ variables that converge to the values $\gamma^t_{d,i}$) it holds that $\frac{1}{C^{t'}_\gamma}\gamma^*_{d,i} \leq \gamma^{t'}_{d,i} \leq C^{t'}_\gamma \gamma^*_{d,i}$. Then, as long as $C^{t'}_\gamma$ is too big, compared to C^t_β, one can show that $C^{t'}_\gamma$ is decreasing (to $C^{t'+1}_\gamma = (C^t_\gamma)^{1/2}$, say), using exactly the same argument we had before. Furthermore, the number of such iterations needed will clearly be logarithmic.

But the same argument as above proves the incomplete tEM updates work as well. Namely, even if we perform only one update of the γ variables, they are guaranteed to improve.

2.3.2 Provable guarantees for initialization

For completeness, we also give here a fairly easy, efficient initialization algorithm. Recall, the goal of this phase is to recover the supports - i.e. to find out which topics are present in a document, and identify the support of each topic. To reiterate the theorem statement:

Theorem 7. If the number of documents is $\Omega(k^4 \log^2 k)$, there is a polynomial-time procedure which with probability $1 - \Omega(\frac{1}{k})$ correctly identifies the supports of the $\beta^*_{i,j}$ and $\gamma^*_{d,i}$ variables.

We will find the topic supports first. Roughly speaking, we will devise a test, which will take as input two documents d, d', and will try to determine if the two documents have a topic in common or not. The test will have no false positives, i.e. will never say NO, if the documents do have a topic in common, but might say NO even if they do. We will then, ensure that with high probability, for each topic we find a pair of documents intersecting in that topic, such that the test says YES.

We will also be able to identify which pairs intersect in exactly one topic, and from this we will be able to find all the topic supports. Having done all of this, finding the topics in each document will be easy as well. Roughly speaking, if a document doesn't contain a given topic, it will not contain all of the discriminative words in that document.

We give the algorithm formally as pseudocode Algorithm 4.

Now, let's proceed to analyze the above algorithm, proceeding in a few parts.

Algorithm 4 Initialization

repeat
  Sample a pair of documents (d, d').
  ▷*Test if (d, d') intersect with no false positives:*
  if $\sum_j \min\{f^*_{d,j}, f^*_{d',j}\} \geq \frac{1}{2T}$ **then**
    $S_{d,d'} := \{j, s.t. f^*_{d,j}, f^*_{d',j} > 0\}$
    ▷*"Weed-out" words that are not in the support of the intersection of (d,d')*
    for all documents $d'' \neq \{d, d'\}$ **do**
      if $\sum_j \min\{f^*_{d,j}, f^*_{d'',j}\} \geq \frac{1}{2s}$ and $\sum_j \min\{f^*_{d',j}, f^*_{d'',j}\} \geq \frac{1}{2s}$ **then**
        $S_{d,d'} = S_{d,d'} \cap j$, s.t $f^*_{d'',j} > 0$
      end if
    end for
  end if
until $K^4 \log^2 K$ times
▷*Determine which $S_{a,b}$ correspond to documents intersecting in one topic only)*
if Set $S_{a,b}$ appears less than $D/K^{2.5}$ times, where D is the total number of documents **then**
  Remove $S_{a,b}$.
end if
if Set $S_{a,b}$ can be written as the union of two other sets $S_{c,d}$, $S_{e,f}$, where neither is contained inside the other **then**
  Remove $S_{a,b}$.
end if
if Set $S_{a,b}$ is strictly contained inside $S_{d,d'}$ for some $S_{d,d'}$ **then**
  Remove $S_{d,d'}$.
end if
Remove duplicates.
The remaining lists $S_{a,b}$ are declared to be topic supports.

2.3.2.1 Constructing a no-false-positives test

First, we describe how one determines the supports of the topics. Let's define $Test(d, d')$ = YES, if $\sum_j \min\{f^*_{d,j}, f^*_{d',j}\} \geq \frac{1}{2T}$, and NO otherwise. Then, we claim the following.

Lemma 18. *If d, d' both contain a topic i_0, s.t. $\gamma^*_{d,i_0} \geq 1/s$, $\gamma^*_{d',i_0} \geq 1/s$ then $Test(d, d')$ = YES. If d, d' do not contain a topic i_0 in common, then $Test(d, d')$ = NO.*

Proof. Let's prove the first claim.

$$\sum_j \min\{\tilde{f}_{d,j}, \tilde{f}_{d',j}\} \geq \sum_j (1-\epsilon) \min\{\beta^*_{i_0,j}\gamma^*_{d,i_0}, \beta^*_{i_0,j}\gamma^*_{d',i_0}\} \geq$$

$$\sum_j (1-\epsilon) 1/s \beta^*_{i_0,j} \geq 1/2s$$

Now, let's prove the second claim. Let's suppose d, d' contain no topic in common.

Let's fix a topic i_0 that belongs to document d. By the "small discriminative words intersection", we have the following property:

$$\sum_{j \in i_0, j \in i'} \beta^*_{i,j} = o(1)$$

for any other topic $i' \neq i_0$.

Denoting by $T_{outside}$ the words belonging to topic i_0, and no topic in document d', and T_{inside} the words belonging to at least one other topic in d', we have

$$\sum_{j \in T_{inside}} \beta^*_{i,j} \leq s \cdot o(1) = o(1)$$

For the words $j \in T_{outside}$, $\min\{f^*_{d,j}, f^*_{d',j}\} = 0$

By the above,

$$\sum_j \min\{\tilde{f}_{d,j}, \tilde{f}_{d',j}\} \leq (1+\epsilon)s^2 o(1) = o(1)$$

Thus, the test will say NO, as we wanted.

□

2.3.2.2 Finding the topic supports from identifying pairs

Let's call d, d' an *identifying pair* of documents for topic i, if d, d' intersect in topic i only, and furthermore the test says YES on that pair.

From this identifying pair, we show how to find the support of the topic i in the intersection. What we'd like to do is just declare the words j, s.t. $f^*_{d,j}, f^*_{d',j}$ are both non-zero as the support of topic i. Unfortunately, this doesn't quite work. The reason is that one might find words j, s.t. they belong to one topic i' in d, and another topic i'' in d''. Fortunately, this is easy to remedy. As per the pseudo-code above, let's call the following operation $WEEDOUT(d, d')$:

- Set $S = \{j, s.t. f^*_{d,j} > 0, f^*_{d',j} > 0\}$.
- For all d'', s.t. $Test(d, d'') = YES$, $Test(d', d'') = YES$:
- Set $S = S \cup \{j, s.t. f^*_{d'',j} > 0\}$
- Return S.

Lemma 19. *With probability* $1 - \Omega(\frac{1}{k})$*, for any pair of documents* d, d' *intersecting in one topic,* $WEEDOUT(d, d')$ *is the support of* S.

Proof. For this, we prove two things. First, it's clear that S is initialized in the first line in a way that ensures that it contains all words in the support of topic i. Furthermore, it's clear that at no point in time we will remove a word j from S that is in the support of topic i. Indeed - if $Test(d, d'') = YES$ and $Test(d', d'') = YES$, then by Lemma 18 document d'' must contain topic i. In this case, $f^*_{d'',j} > 0$, and we won't exclude j from S.

So, we only need to show that the words that are not in the support of topic i will get removed.

Let d, d' intersect in a topic i. Let a word j be outside the support of a given topic i. Because of the independent topic inclusion property, the probability that a document d'' contains topic i, and no other topic containing j is $\Omega(1/K)$.

Since the number of documents is $\Omega(k^4 \log^2 k)$, by Chernoff, the probability that there is a document d'', s.t. $Test(d, d'') = YES$, $Test(d', d'') = YES$, but $f^*_{d'',j} = 0$, is $1 - \Omega(\frac{1}{e^{K^2 \log^2 k}})$. Union bounding over all words j, as well as pairs of documents d, d', we get that for any documents d, d' intersection in a topic i, we get the claim we want.

□

2.3.2.3 Finding the identifying pairs

Finally, we show how to actually find the identifying pairs. The main issue we need to handle are documents that do intersect, and the TEST returns yes, but they intersect in more than one topic. There's two ingredients to ensuring this is true in the above algorithm.

- First, we delete all sets in the list of sets $S_{a,b}$ that show up less than $D^2/k^{2.5}$ number of times.
- Second, we remove sets that can be written as the union of two other sets $S_{c,d}$, $S_{e,f}$, where neither of the two is contained inside the other.
- After this, we delete the non-maximal sets in the list.

The following lemma holds:

Lemma 20. *Each topic has $\Omega(D^2/k^2)$ identifying pairs with probability $1 - \Omega(\frac{1}{k})$.*

Proof. Let $\mathcal{I}_i$ be the event that there are at least $\Omega(D^2/k^2)$ identifying pairs for topic i. Let R_i be a random variable denoting the number of documents which have topic i as a dominating topic. Furthermore, let $\mathcal{M}_i$ be the event that there are at least $\frac{R_i^2}{2} - k\sqrt{R_i^2}$ identifying pairs among the R_i ones that have i as a dominating topic. By the dominant topic equidistribution property, probability that a document d has a topic i as a dominating topic is at least C/k for some constant C. Then, clearly,

$$\Pr[\cap_{i=1}^k \mathcal{I}_i] \geq \Pr\left[\cap_{i=1}^k \left(R_i \geq \frac{1}{2} C \frac{D}{k}\right)\right] \Pr\left[\cap_{i=1}^k \mathcal{M}_i | \cap_{i=1}^k \left(R_i \geq \frac{1}{2} C \frac{D}{k}\right)\right]$$

Let's estimate $\Pr\left[\cap_{i=1}^k \left(R_i \geq \frac{1}{2} C \frac{D}{k}\right)\right]$ first. The probabilities that different documents have i_0 as the dominating topic are clearly independent, so by Chernoff, if R_i is the number of documents where i is the dominating topic,

$$\Pr[R_i \geq (1-\epsilon) C \frac{D}{k}] \geq 1 - e^{-\frac{\epsilon^2}{3} C \frac{D}{k}}$$

Since $D = \Omega(k^2)$, plugging in $\epsilon = \frac{1}{2}$, $\Pr[R_i < \frac{1}{2} C \frac{D}{k}] \geq 1 - e^{-\Omega(k)}$. Union bounding over all topics, we get that with probability $\Pr\left[\cap_{i=1}^k \left(R_i \geq \frac{1}{2} C \frac{D}{k}\right)\right] \geq 1 - \frac{1}{k}$.

Now, let's consider $\Pr\left[\cap_{i=1}^{k}\mathcal{M}_i|\cap_{i=1}^{k}\left(R_i \geq \frac{1}{2}C\frac{D}{k}\right)\right]$. The event $\cap_{i=1}^{k}\left(R_i \geq \frac{1}{2}C\frac{D}{k}\right)$ can be written as the disjoint union of events

$$\{\mathbb{D} = \cup_{i=1}^{k} D_i, \forall i \neq j, D_i \cap D_j = \emptyset\}$$

where $\mathbb{D}$ is the set of all documents, D_i is the set of documents that have i as the dominating topic, and $|D_i| \geq \frac{1}{2}C\frac{D}{k}, \forall i$. (i.e. all the partitions of $\mathbb{D}$ into k sets of sufficiently large size). Evidently, if we prove a lower bound on $\Pr\left[\cap_{i=1}^{k}\mathcal{M}_i|E\right]$ for any such event E, it will imply a lower bound on $\Pr\left[\cap_{i=1}^{k}\mathcal{M}_i|\cap_{i=1}^{k}\left(R_i \geq \frac{1}{2}C\frac{D}{k}\right)\right]$. For any such event, consider two documents $d, d' \in \{D_i\}$, i.e. having i as the dominating topic. Let $\mathcal{I}_{d,d'}$ be an indicator variable denoting the event that d, d' do not intersect in an additional topic. $\Pr[\mathcal{I}_{d,d'} = 1] = 1 - o(1)$, by the independent topic inclusion property and the events $\mathcal{I}_{d,d'}$ are easily seen to be pairwise independent. Furthermore, $\mathrm{Var}[\mathcal{I}_{d,d'}] = o(1)$. By Chebyshev's inequality,

$$\Pr\left[\sum_{d,d'\in D_i}\mathcal{I}_{d,d'} \geq \frac{1}{2}D_i^2 - c\sqrt{D_i^2}\right] \geq 1 - \frac{1}{c^2}$$

If $R_i = \Omega(k\log k)$, plugging in $c = K$, we get that $\Pr\left[\sum_{d,d'\in D_i}\mathcal{I}_{d,d'} = \Omega(D_i^2)\right] \geq 1 - \Omega(\frac{1}{k^2})$. Hence, $\Pr\left[\cap_{i=1}^{k}\mathcal{M}_i|E\right] \geq 1 - \frac{1}{k}$, by a union bound, which implies $\Pr\left[\cap_{i=1}^{k}\mathcal{M}_i|\cap_{i=1}^{k}\left(N_i \geq \frac{1}{2}C\frac{D}{k}\right)\right] \geq 1 - \frac{1}{k}$.

Putting all of the above together, if $D = \Omega(k^2\log k)$, with probability $1 - \Omega(\frac{1}{k})$, all topics have $\Omega(D^2/k^2)$ identifying pairs, which is what we want.

□

The lemma implies that with probability $1 - \Omega(\frac{1}{k})$, we will not eliminate the sets $S_{a,b}$ corresponding to topic supports.

We introduce the following concept of a "configuration". A set of words C will be called a "configuration" if it can be constructed as the intersection of the discriminative words in some set of topics, i.e.

Definition. A set of words C is called a configuration if there exists a set $I = \{I_1, \ldots, I_{|I|}\}$ of topics, s.t.

$$C = \cap_{i=1}^{|I|} W_{I_i}$$

Let's call the minimal size of a set I that can produce C the generator size of C.

Now, we claim the following fact:

Lemma 21. *If a configuration C has generator size ≥ 3, then with probability $1 - \Omega(\frac{1}{k})$, it cannot appear as one of the sets $S_{a,b}$ after step 2 in the WEEDOUT procedure.*

Proof. Since C has generator size at least 3, if two sets d, d' intersect in less than two topics, then step 1 in WEEDOUT

cannot produce $S_{a,b}$ which is equal to C. Hence, prior to step 2, C can only appear as $S_{d,d'}$ for d, d' that intersect in at least 3 topics.

Let $\mathcal{I}_{d,d'}$ be an indicator variable denoting the fact that the pair of documents d, d' intersects in at least 3 topics. We have $\Pr[\mathcal{I}_{d,d'} = 1] \leq 1/k^3 + 1/k^4 + \ldots 1/k^T = O(1/k^3)$ by the independent topic inclusion property.

If $\mathcal{I}_3$ is a variable denoting the total number of documents that intersect in at least 3 topics, again by Chebyshev as in Lemma 20 we get:

$$\Pr[\mathcal{I}_3 \geq \Theta(D/k^3) - c\Theta(\sqrt{D}/k^{3/2})] \geq 1 - \frac{1}{c^2}$$

Again, by putting $c = \sqrt{k}$, since the number of documents is $k^4 \log^2 k$, with probability $1 - \frac{1}{k}$, all configurations with generator size ≥ 3 cannot appear as one of the sets $S_{a,b}$, as we wanted.

□

This means that after the WEEDOUT step, with probability $1 - \Omega(\frac{1}{k})$, we will just have sets $S_{a,b}$ corresponding to configurations generated by two topics or less. The options for these are severely limited: they have to be either a topic support, the union of two topic supports, or the intersection of two topic supports. We can handle this case fairly easily, as proven in the following lemma:

Lemma 22. *After the end of step 3, with probability* $1 - \Omega(\frac{1}{k})$*, the only remaining* $S_{a,b}$ *are those corresponding to topic supports.*

Proof. First, when we check if some $S_{d,d'}$ is the union of two other sets and delete it if yes, I claim we will delete the sets equal to configurations that correspond to unions of two topic supports (and nothing else). This is not that difficult to see: certainly the sets that do correspond to configurations of this type will get deleted.

On the other hand, if it's the case that $S_{a,b}$ corresponds to a single topic support, we won't be able to write it as the union of two sets $S_{d,d'}$, $S_{d'',d'''}$, unless one is contained inside the other - this is ensured by the existence of discriminative words.

Hence, after the first two passes, we will only be left with sets that are either topic supports, or intersections of two topic supports. Then, removing the non-maximal is easily seen to remove the sets that are intersections, again due to the existence of discriminative words.

□

2.3.2.4 Finding the document supports

Now, given the supports of each topic, for each document, we want to determine the topics which are non-zero in it. The algorithm is given in 5:

Algorithm 5 Finding document supports

Initialize $R = \emptyset$.
for each i **do**
  Compute Score(i) = $\sum_{j \in Support(i) \setminus R} \tilde{f}_{d,j}$
end for
Find i^* such that Score(i^*) is maximum.
while Score(i^*) > 0 **do**
  Output i^* to be in the support of d.
  $R = R \cup support(i^*)$
  Recompute Score for every other topic.
  Find i^* with maximum score.
end while

Lemma 23. *If a topic i_0 is such that $\gamma^*_{d,i_0} > 0$, it will be declared as "IN". If a topic i_0 is such that $\gamma^*_{d,i_0} = 0$, it will be declared as out.*

Proof. Consider a topic i. At any iteration of the while cycle, consider $\sum_{j \in Support(i) \setminus R} \tilde{f}_{d,j}$. Clearly, $\tilde{f}_{d,j} \geq (1-\epsilon)\gamma^*_{d,i}\beta^*_{i,j}$. Also $\sum_{j \in R} \beta^*_{i,j} = so(1)$. Hence,

$$\sum_{j \in Support(i) \setminus R} \tilde{f}_{d,j} \geq (1-\epsilon)\gamma^*_{d,i}(1 - so(1)) \geq \frac{1}{2}\gamma^*_{d,i}$$

So, topic i will be added eventually.

On the other hand, let's assume the document doesn't contain a given topic i_0. Let's call B the set of words j which are in the support of i_0, and belong to at least one of the topics in document d. Then, $\sum_{j \in i_0} \tilde{f}_{d,j} = \sum_{j \in B} \tilde{f}_{d,j}$. Let i^* be the topic which is present in the document but not added yet and has maximum value of $\gamma^*_{d,i}$. Then

$$\sum_{j \in B} \tilde{f}_{d,j} \leq (1+\epsilon) \sum_{i \in d} \sum_{j \in B} \gamma^*_{d,i}\beta^*_{i,j} \leq$$

$$(1+\epsilon)\gamma^*_{d,i^*} \sum_{i \in d} \sum_{j \in B} \beta^*_{i,j} \leq$$

$$(1+\epsilon)s\gamma^*_{d,i^*} o(1) \leq \gamma^*_{d,i^*}] \cdot o(1)$$

Hence, topic i^* will always get preference over i_0. Once all the topics which are present in the document have been added, it is clear that no more topic will be added since score will be 0.

□

This finally finishes the proof of Theorem 7.

2.4 Case study 2: Dominating topics, seeded initialization

Next, we'll consider an initialization which is essentially what the current implementation of LDA-c uses. Namely, we will call the following initialization a *seeded* initialization:

- For each topic i, the user supplies a document d, in which $\gamma^*_{d,i} \geq C_l$.
- We treat the document as if it only contains topic i and initialize with $\beta^0_{i,j} = f^*_{d,j}$.

We show how to modify the previous analysis to show that with a few more assumptions, this strategy works as well. Firstly, we will have to assume anchor words, that make up a decent fraction of the mass of each topic. Second, we also assume that the words have a bounded *dynamic range*, i.e. the values of a word in two different topics are within a constant B from each other. The documents are still gapped, but the gap now must be larger. Finally, in roughly $1/B$ fraction of the documents where topic i is dominant, that topic has proportion $1-\delta$, for some small (but still constant) δ. A similar assumption (a small fraction of almost pure documents) appeared in a recent paper by (Bansal et al., 2014). Formally, we have:

- *Small dynamic range and large fraction of anchors*: For each discriminative words, if $\beta^*_{i,j} \neq 0$ and $\beta^*_{i',j} \neq 0$, $\beta^*_{i,j} \leq B\beta^*_{i',j}$. Furthermore, each topic i has anchor words, such that their total weight is at least p.
- *Gapped documents*: In each document, the largest topic has proportion at least C_l, and all the other topics are at most C_s, s.t.
$$C_l - C_s \geq \frac{1}{p}\left(\sqrt{2\left(p\log(\frac{1}{C_l}) + (1-p)\log(BC_l)\right)} + \sqrt{\log(1+\epsilon)}\right) + \epsilon$$
- *Small fraction of* $1-\delta$ *dominant documents*: Among all the documents where topic i is dominating, in a $8/B$ fraction of them, $\gamma^*_{d,i} \geq 1-\delta$, where
$$\delta := \min\left(\frac{C_l^2}{2B^3} - \frac{1}{p}\left(\sqrt{2\left(p\log(\frac{1}{C_l}) + (1-p)\log(BC_l)\right)} + \sqrt{\log(1+\epsilon)}\right) - \epsilon, 1 - \sqrt{C_l}\right)$$

The dependency between the parameters B, p, C_l is a little difficult to parse, but if one thinks of C_l as $1-\eta$ for η small, and $p \geq 1 - \frac{\eta}{\log B}$, since $\log(\frac{1}{C_l}) \approx 1+\eta$, roughly we want that $C_l - C_s \gg \frac{2}{p}\sqrt{\eta}$. (In other words, the weight we require to have on the anchors depends only *logarithmically* on the range B.) In the documents where the dominant topic has proportion $1-\delta$, a similar reasoning as above gives that we want is approximately $\gamma^*_{d,i} \geq 1 - \frac{1-2\eta}{2B^3} + \frac{2}{p}\sqrt{\eta}$. The precise statement is as follows:

Theorem 24 ((Awasthi and Risteski, 2015))**.** *Given an instance of topic modelling satisfying the properties specified above, where the number of documents is* $\Omega(\frac{k\log^2 n}{\epsilon^2})$*, if we initialize with seeded initialization, after* $O\left(\log(1/\epsilon') + \log n\right)$

of KL-tEM updates, we recover the topic-word matrix and topic proportions to multiplicative accuracy $1 + \epsilon'$*, if* $1 + \epsilon' \geq \frac{1}{(1-\epsilon)^7}$.

The proof is carried out in a few phases:

- *Phase I: Anchor identification*: We show that as long as we can identify the dominating topic in each of the documents, anchor words will make progress: after $O(\log n)$ number of rounds, the values for the topic-word estimates will be almost zero for the topics for which word w is not an anchor. For topic for which a word is an anchor we'll have a good estimate.

- *Phase II: Discriminative word identification*: After the anchor words are properly identified in the previous phase, if $\beta^*_{i,j} = 0$, $\beta^t_{i,j}$ will keep dropping and quickly reach almost zero. The values corresponding to $\beta^*_{i,j} \neq 0$ will be decently estimated.

- *Phase III: Alternating minimization*: After Phase I and II above, we are back to the scenario of the previous section: namely, there is improvement in each next round.

During Phase I and II the intuition is the following: due to our initialization, even in the beginning, each topic is "correlated" with the correct values. In a γ update, we are minimizing $KL(\tilde{f}_d \| f_d)$ with respect to the γ_d variables, so we need a way to argue that whenever the β estimates are not too bad, minimizing this quantity provides an estimate about how far the optimal γ_d variables are from γ^*_d. We show the following useful claim:

Lemma 25. *If, for all topics* i*,* $KL(\beta^*_i \| \beta^t_i) \leq R_\beta$*, and* $min_{\gamma_d \in \Delta_K} KL(\tilde{f}_{d,j} \| f_{d,j}) \leq R_f$*, after running a KL divergence minimization step with respect to the* γ_d *variables, we get that* $\|\gamma^*_d - \gamma_d\|_1 \leq \frac{1}{p}(\sqrt{\frac{1}{2}R_\beta} + \frac{1}{2}\sqrt{R_f}) + \epsilon$.

This lemma critically uses the existence of anchor words - namely we show $\|\beta^* v\|_1 \geq p\|v\|_1$. Intuitively, if one thinks of v as $\gamma^* - \gamma^t$, $\|\beta^* v\|_1$ will be large if $\|v\|_1$ is large. Hence, if $\|\beta^* - \beta^t\|_1$ is not too large, whenever $\|f^* - f^t\|_1$ is small, so is $\|\gamma^* - \gamma^t\|_1$. We will be able to maintain R_β and R_f small enough throughout the iterations, so that we can identify the largest topic in each of the documents.

We proceed to making these ideas more formal in the coming sections. The proof will be in a few phases again:

- *Phase I: Anchor identification*: First, we will show that as long as we can identify the dominating topic in each of the documents, the anchor words will make progress, in the sense that after $O(\log N)$ number of rounds, the values for the topic-word estimates will be almost zero for the topics for which the word is not an anchor, and lower bounded for the one for which it is.

- *Phase II: Discriminative word identification*: Next, we show that as long as we can identify the dominating topics in each of the documents, and the anchor words were properly identified in the previous phase, the values

of the topic-word matrix for words which do not belong to a certain topic will keep dropping until they reach almost zero, while being lower bounded for the words that do.

- For Phase I and II above, we will need to show that the dominating topic can be identified at any step. Here we'll leverage the fact that the dominating topic is sufficiently large, as well as the fact that the anchor words have quite a large weight.

- *Phase III: Alternating minimization*: Finally, we show that after Phase I and II above, we are back to the scenario of the previous section: namely, there is a "boosting" type of improvement in each next round.

2.4.1 Estimates on the dominating topic

Before diving into the specifics of the phases above, we will show what the conditions we need are to be able to identify the dominating topic in each of the documents. For notational convenience, let Δ_m be the m-dimensional simplex: $x \in \Delta_m$ iff $\forall i \in [m], 0 \leq x_i \leq 1$ and $\sum_i x_i = 1$.

First, during a γ update, we are minimizing $KL(\tilde{f}_d || f_d)$ with respect to the γ_d variables, so we need some way or arguing that whenever the β estimates are not too bad, minimizing this quantity also quantifies how far the γ_d variables are from γ_d^*.

Formally, we'll show the following:

Lemma 26. *If, for all i, $KL(\beta_i^* || \beta_i^t) \leq R_\beta$, and $\min_{\gamma_d \in \Delta_K} KL(\tilde{f}_d || f_d) \leq R_f$, after running a KL divergence minimization step with respect to the γ_d variables, we get that $\|\gamma_d^* - \gamma_d\|_1 \leq \frac{1}{p}(\sqrt{\frac{1}{2}R_\beta} + \sqrt{\frac{1}{2}R_f}) + \epsilon$.*

We will start with the following simple helper claim:

Lemma 27. *If the word-topic matrix β is such that in each topic the anchor words have total probability at least p, then $\|\beta^* v\|_1 \geq p\|v\|_1$.*

Proof.

$$\|\beta^* v\|_1 = \sum_j |\sum_i \beta_{i,j}^* v_i| \geq \sum_i \sum_{j \in W_i} |\beta_{i,j}^* v_i| \geq \sum_i p|v_i| \geq p\|v\|_1$$

□

Lemma 28. *If, for all i, $KL(\beta_i^* || \beta_i^t) \leq R_\beta$, and $\min_{\gamma_d \in \Delta_K} KL(\tilde{f}_d || f_d) \leq R_f$, after running a KL divergence minimization step with respect to the γ_d variables, we get that $\|\gamma_d^* - \gamma_d\|_1 \leq \frac{1}{p}(\sqrt{\frac{1}{2}R_\beta} + \sqrt{\frac{1}{2}R_f}) + \epsilon$.*

Proof. First, observe that $\min_{\gamma_d \in \Delta_K} KL(\tilde{f}_d || f_d) \leq R_f$, at the the optimal γ_d, we have that $\|\tilde{f}_d - f_d\|_1^2 \leq \frac{1}{2}R_f$, i.e. $\|\tilde{f}_d - f_d\| \leq \sqrt{\frac{1}{2}R_f}$, by Pinsker's inequality.

We will show that if $\|\gamma_d^* - \gamma_d\|_1$ is large, so must be $\|\tilde{f}_d - f_d\|_1$, and hence $KL(\tilde{f}_d \| f_d)$ - which will contradict the above upper bound.

Let's consider β^* as N by K matrix, and γ^* and f^* as K-dimensional vectors. Let $\beta^*\gamma^*$ just denote matrix-vector multiplication - so $f^* = \beta^*\gamma^*$. For any other vector $\hat{\gamma}$, let's denote $\hat{f} = \beta^t\tilde{\gamma}$. Then:

$$\|\tilde{f} - \hat{f}\|_1 = \|\tilde{f} - \beta^t\tilde{\gamma}\|_1 = \|\tilde{f} - (\beta^* + (\beta^t - \beta^*))\tilde{\gamma}\|_1 \geq$$

$$\|\tilde{f} - \beta^*\tilde{\gamma}\|_1 - \|(\beta^t - \beta^*)\tilde{\gamma}\|_1 \tag{2.4.1}$$

Hence, $\|\tilde{f} - \beta^*\tilde{\gamma}\|_1 \leq \|(\beta^t - \beta^*)\tilde{\gamma}\|_1 + \|\tilde{f} - \hat{f}\|_1$. However,

However,

$$\|(\beta^t - \beta^*)\gamma\|_1 \leq \max_i \sum_j |\beta^t_{i,j} - \beta^*_{i,j}| \leq \max_i \sqrt{\frac{1}{2}KL(\beta^*_i \| \beta^t_i)} \leq \sqrt{\frac{1}{2}R_\beta} \tag{2.4.2}$$

The first inequality is a property of induced matrix norms, the second is via Pinsker's inequality.

So, by 2.4.1 and 2.4.2, $\|\tilde{f} - \beta^*\tilde{\gamma}\|_1 \leq \sqrt{\frac{1}{2}R_\beta} + \sqrt{\frac{1}{2}R_f}$. But now, finally, Lemma 27 implies that $\|\gamma_d^* - \gamma_d\|_1 \leq \frac{1}{p}(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}) + \epsilon$.

□

Lemma 29. *Suppose that for the dominating topic i in a document d, $\gamma^*_{d,i} \geq C_l$, and for all other topics i', $\gamma^*_{d,i'} \leq C_s$, s.t. $C_l - C_s > \frac{1}{p}(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}) + \epsilon$. Then, the above test identifies the largest topic. Furthermore, $\frac{1}{2}\gamma^*_{d,i} \leq \gamma^t_{d,i} \leq \frac{3}{2}\gamma^*_{d,i}$*

Proof. By Lemma 28, and the relationship between l_1 and total variation distance between distributions, we have that $|\gamma^t_{d,i} - \gamma^*_{d,i}| \leq \frac{1}{2}\left(\frac{1}{p}\left(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}\right) + \epsilon\right)$.

For the dominating topic i, $\gamma^t_{d,i} \geq C_l - \frac{1}{2}\left(\frac{1}{p}\left(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}\right) + \epsilon\right)$. On the other hand, for any other topic i', $\gamma^t_{d,i'} \leq C_s + \frac{1}{2}\left(\frac{1}{p}\left(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}\right) + \epsilon\right)$. Since $C_l - C_s \geq \frac{1}{p}\left(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}\right) + \epsilon$, $\gamma^t_{d,i} > \gamma^t_{d,i'}$, so the test works.

On the other hand, since $\gamma^t_{d,i} \geq \gamma^*_{d,i} - \left(\frac{1}{p}\left(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}\right) + \epsilon\right) \geq \gamma^*_{d,i} - \frac{1}{2}\gamma^*_{d,i} = \frac{1}{2}\gamma^*_{d,i}$. Similarly, $\gamma^t_{d,i} \leq \gamma^*_{d,i} + \frac{1}{p}\left(\sqrt{\frac{1}{2}R_f} + \sqrt{\frac{1}{2}R_\beta}\right) + \epsilon \leq \gamma^*_{d,i} + \frac{1}{2}\gamma^*_{d,i} = \frac{3}{2}\gamma^*_{d,i}$.

□

2.4.2 Phase I: Determining the anchor words

We proceed as outlined. In this section we show that in the first phase of the algorithm, the anchor words will be identified - by this we mean that we will be able to show that if a word j is an anchor for topic i, $\beta^t_{i,j}$ will be within a factor of roughly 2 from $\beta^*_{i,j}$, and $\beta^t_{i',j}$ will be almost 0 for any other topic i'.

We will assume throughout this and the next section that we can identify what the dominating topic is, and that we have an estimate of the proportion of the dominating topic to within a factor of 2. (We won't restate this assumption in all the lemmas in favor of readability.)

We will return to this issue after we've proven the claims of Phases I and II modulo this claim.

The outline is the following. We show that at any point in time, by virtue of the initialization, $\beta^t_{i,j}$ is pretty well lower bounded (more precisely it's at least constant times $\beta^*_{i,j}$). This enables us to show that $\beta^t_{i',j}$ will halve at each iteration - so in some polynomial number of iterations will be basically 0.

2.4.2.1 Lower bounds on the $\beta^t_{i,j}$ values

We proceed as outlined above. We show here that the $\beta^t_{i,j}$ variables are lower bounded at any point in time. More precisely, we show the following lemma:

Lemma 30. *Let j be an anchor word for topic i, and let $i' \neq i$. Suppose that $\beta^t_{i',j} \leq \beta^t_{i,j}$. Then, $\beta^{t+1}_{i,j} \geq (1-\epsilon)C_l\beta^*_{i,j}$ holds.*

Proof. We'll prove a lower bound on each of the terms $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j}$. Since the update on the β variables is a convex combination of terms of this type, this will imply a lower bound on $\beta^{t+1}_{i,j}$.

For this, we upper bound $f^t_{d,j}$. We have:

$$f^t_{d,j} = \beta^t_{i,j}\gamma^t_{d,i} + \sum_{i' \neq i} \beta^t_{i',j}\gamma^t_{d,i'}$$

This means that $f^t_{d,j}$ is a convex combination of terms, each of which is at most $\beta^t_{i,j}$. Hence, $f^t_{d,j} \leq \beta^t_{i,j}$ holds. But then $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} \geq \tilde{f}_{d,j} \geq (1-\epsilon)\beta^*_{i,j}\gamma^*_{d,i} \geq (1-\epsilon)C_l\beta^*_{i,j}$. This implies $\beta^{t+1}_{i,j} \geq (1-\epsilon)C_l\beta^*_{i,j}$, as we wanted.

□

2.4.2.2 Decreasing $\beta^t_{i',j}$ values

We'll bootstrap to the above result. Namely, we'll prove that whenever $\beta^t_{i,j} \geq 1/C_\beta\beta^*_{i,j}$ for some constant C_β, the $\beta^t_{i',j}$ values decrease multiplicatively at each round. Prior to doing that, the following lemma is useful. It will state that whenever the values of the variables $\beta^t_{i',j}$ are somewhat small, we can get some reasonable lower bound on the values $\gamma^t_{d,i}$ we get after a step of KL minimization with respect to the γ variables.

Lemma 31. *Let j be an anchor for topic i, and let $i' \neq i$. Let $\beta^t_{i',j} \leq b\beta^t_{i,j}$. Then, for any document d, when performing KL divergence minimization with respect to the variables γ_d, for the optimum value $\gamma^t_{d,i}$, it holds that $\gamma^t_{d,i} \geq (1-\epsilon)\frac{p}{1-b}\gamma^*_{d,i} - \frac{b}{1-b}$.*

Proof. The KKT conditions 2.3.1 imply that if we denote A_i the set of anchors in topic i, $\sum_{j\in A_i} \frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} \le 1$. By the assumption of the lemma,

$$f^t_{d,j} \le b^t_{i,j}\gamma^t_{d,i} + bb^t_{i,j}(1-\gamma^t_{d,i})$$

Since $\tilde{f}_{d,j} \ge (1-\epsilon)\beta^*_{i,j}\gamma^*_{d,i}$, this implies $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} \ge (1-\epsilon)\beta^*_{i,j}\frac{\gamma^*_{d,i}}{\gamma^t_{d,i}(1-b)+b}$, i.e. $\sum_{j\in A_i}(1-\epsilon)\beta^*_{i,j}\frac{\gamma^*_{d,i}}{\gamma^t_{d,i}(1-b)+b} \le 1$. Rearranging the terms, we get

$$\gamma^t_{d,i} \ge (1-\epsilon)\sum_{j\in A_i}\beta^*_{i,j}\frac{\gamma^*_{d,i}}{1-b} - \frac{b}{1-b} \ge (1-\epsilon)p\gamma^*_{d,i} - \frac{b}{1-b}$$

as we needed.

□

With this in place, we show that the value $\beta^t_{i',j}$ when j is an anchor for topic $i \neq i'$, decreases by a factor of 2 after the update for the β variables.

This requires one more new idea. Intuitively, if we view the update as setting $\beta^{t+1}_{i',j}$ to $\beta^t_{i,j}$ multiplied by a convex combination of terms $\frac{f^*_{d,j}}{f^t_{d,j}}$, a large number of them will be zero, just because $f^*_{d,j} = 0$ unless topic i belongs to document d.

By the topic equidistribution property then, the probability that this happens is only $O(1/k)$, so if the weight in the convex combination on these terms is reasonable, we will multiply $\beta^t_{i,j}$ by something less than 1, which is what we need.

Lemma 31 says that if $\gamma^*_{d,i}$ is reasonably large, we will estimate it somewhat decently. If $\gamma^*_{d,i}$ is small, then $f^*_{d,j}$ would be small anyway.

So we proceed according to this idea.

Lemma 32. *Let j be an anchor for topic i. Let $\beta^t_{i',j} \le b\beta^t_{i,j}$ for $i' \neq i$, and let $\beta^t_{i,j} \ge 1/C_\beta\beta^*_{i,j}$ for some constant C_β. Then, $\beta^{t+1}_{i',j} \le b/2\beta^*_{i,j}$*

Proof. We will split the β update as

$$\beta^{t+1}_{i',j} = \beta^t_{i',j}\left(\frac{\sum_{d\in D_1}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}} + \frac{\sum_{d\in D_2}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}} + \frac{\sum_{d\in D_3}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}}\right)$$

for some appropriately chosen partition of the documents into three groups D_1, D_2, D_3.

Let D_1 be documents which do not contain topic i at all, D_2 documents which do contain topic i, and $\gamma^*_{d,i} \ge \frac{2b}{p}$, and D_3 documents which do contain topic i and $\gamma^*_{d,i} < \frac{2b}{p}$.

The first part will just vanish because word j is an anchord word for topic i, and topic i does not appear in it, so $f^*_{d,j} = 0$ for all documents $d \in D_1$.

The second summand we will upper bound as follows. First, we upper bound $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}$. We have that $f^t_{d,j} \geq \beta^t_{i,j}\gamma^t_{d,i} \geq 1/C_\beta \beta^*_{i,j}\gamma^t_{d,i}$. However, we can use Lemma 31 to lower bound $\gamma^t_{d,i}$. We have that $\gamma^t_{d,i} \geq (1-\epsilon)(\frac{p}{1-b}\gamma^*_{d,i} - \frac{b}{1-b}) \geq (1-\epsilon)\frac{p}{2(1-b)}\gamma^*_{d,i}$. This alltogether implies $\frac{\tilde{f}_{d,j}}{f^t_{d,j}} \leq \frac{1}{1-\epsilon}\frac{2(1-b)C_\beta}{p}$. Hence,

$$\beta^t_{i',j}\frac{\sum_{d\in D_2}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}} \leq \frac{1}{1-\epsilon}\frac{2C_\beta}{p}(1-b)\beta^t_{i',j}\frac{\sum_{d\in D_2}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}}$$

Furthermore, $\sum_d \gamma^t_{d,i'} \geq \frac{1}{2}|D|C_l$. On the other hand, we claim $\sum_{d\in D_2}\gamma^t_{d,i'} = O(k/|D|)$. Recall that D is the set of documents where topic i' is the dominating topic - so by definition they contain topic i. On the other hand, if a document is in D_2 then it contains topic i as well. However, by the independent topic inclusion property, the probability that a document with dominating topic i' contains topic i as well is $O(1/k)$. Hence,

$$\beta^t_{i',j}\frac{\sum_{d\in D_2}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}} = O(\frac{1}{k})b\beta^t_{i,j}$$

For the third summand we provide a trivial bound for the terms $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i',j}\gamma^t_{d,i'}$:

$$\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i',j}\gamma^t_{d,i'} \leq (1+\epsilon)\beta^*_{i,j}\gamma^*_{d,i} \leq (1+\epsilon)\beta^*_{i,j}\frac{2b}{p}$$

Since again, $\sum_d \gamma^t_{d,i'} \geq \frac{1}{2}|D|C_l$, and again, the number of document in D_3 is at most $O(1/k)$ for the same reasons as before, we have that

$$\beta^t_{i',j}\frac{\sum_{d\in D_3}\frac{f^*_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}} \leq O(1/k)b\beta^*_{i,j} = O(1/k)b\beta^t_{i,j}$$

since $\beta^t_{i,j} \geq \frac{1}{C_\beta}\beta^*_{i,j}$.

From the above three bounds, we get that $\beta^{t+1}_{i',j} \leq O(1/k)b\beta^t_{i,j} \leq \frac{b}{2}\beta^t_{i,j}$.

□

Now, we just have to put together the previous two claims: namely we need to show that the conditions for the decay of the non-anchor topic values, and the lower bound on the anchor-topic values are actually preserved during the iterations. We will hence show the following:

Lemma 33. *Suppose we initialize with seeded initialization. Then, after t rounds, if j is an anchor word for topic i,* $\beta^t_{i,j} \geq (1-\epsilon)C_l\beta^*_{i,j}$, *and* $\beta^t_{i',j} \leq 2^{-t}C_s\beta^*_{i,j}$.

Proof. We prove this by induction.

Let's cover the base case first. In the seed document corresponding to topic i, $\gamma^*_{d,i} \geq C_l$, so at initialization $\beta^0_{i,j} \geq$

$C_l\beta^*_{i,j}$. On the other hand, if topic i appears in the seed document for topic i', then after initialization $\beta^0_{i',j} \leq C_s\beta^*_{i,j} < \beta^0_{i,j}$. Hence, at initialization, the claim is true.

On to the induction step. If the claim were true at time step t, since $\beta^t_{i',j} \leq 2^{-t}C_s\beta^*_{i,j}$, by Lemma 30, $\beta^{t+1}_{i,j} \geq C_l\beta^*_{i,j}$ - so the lower bound still holds at time $t+1$. On the other hand, since $\beta^t_{i,j} \geq C_l\beta^*_{i,j}$, by Lemma 32, at time $t+1$, $\beta^t_{i',j} \leq 2^{-(t+1)}C_s\beta^*_{i,j}$.

Hence, the claim we want follows.

□

Finally, we show the easy lemma that after the values $\beta^t_{i',j}$ have decreased to (almost) 0, $\beta^t_{i,j} \geq \frac{1}{2}\beta^*_{i,j}$.

Lemma 34. *Let word j be an anchor word for topic i. Suppose $\beta^t_{i',j} \leq 2^{-t}C_s\beta^*_{i,j}$ and*

$$t > 10 \ \max(\log(n), \log(\frac{1}{\gamma^*_{\min}}), \log(\frac{1}{\beta^*_{\min}}))$$

Then $4\beta^*_{i,j} \geq \beta^{t+1}_{i,j} \geq \frac{1}{4}\beta^*_{i,j}$.

Proof. Let us do the lower bound first. It's easy to see $\sum_{i'} \beta^t_{i',j}\gamma_{d,i'} \leq 2\beta^t_{i,j}\gamma^t_{d,i}$. Hence,

$$\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j}\gamma^t_{d,i} = \frac{\tilde{f}_{d,j}}{\sum_{i'}\beta^t_{i',j}\gamma^t_{d,i'}}\beta^t_{i,j}\gamma^t_{d,i} \geq$$

$$\frac{1}{2}\frac{\tilde{f}_{d,j}}{\beta^t_{i,j}\gamma^t_{d,i}}\beta^t_{i,j}\gamma^t_{d,i} \geq (1-\epsilon)\frac{1}{2}\beta^*_{i,j}\gamma^*_{d,i}$$

Hence, after the update,

$$\beta^{t+1}_{i,j} \geq (1-\epsilon)\frac{1}{2}\beta^*_{i,j}\frac{\sum_d \gamma^*_{d,i}}{\sum_d \gamma^t_{d,i}} \geq \frac{1}{4}\beta^*_{i,j}$$

since $\gamma^t_{d,i} \leq 2\gamma^*_{d,i}$.

The upper bound is similar. Since $\sum_{i'} \beta^t_{i',j}\gamma_{d,i'} \geq \beta^t_{i,j}\gamma^t_{d,i}$,

$$\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j}\gamma^t_{d,i} \leq \tilde{f}_{d,j} \leq (1+\epsilon)\beta^*_{i,j}\gamma^*_{d,i}$$

Hence,

$$\beta^{t+1}_{i,j} \leq (1+\epsilon)\beta^*_{i,j}\frac{\sum_d \gamma^*_{d,i}}{\sum_d \gamma^t_{d,i}} \leq 2\beta^*_{i,j}$$

since $\gamma^t_{d,i} \geq \frac{1}{2}\gamma^*_{d,i}$. This certainly implies the claim we want.

□

Furthermore, the following simple application of Lemma 31 is immediate and useful:

Lemma 35. *Let* $t > 10\max(\log n, \log \frac{1}{\gamma^*_{\min}}, \log \frac{1}{\beta^*_{\min}})$. *Then,* $\gamma^t_{d,i} \geq \frac{p}{2}\gamma^*_{d,i}$.

2.4.3 Discriminative words

We established in the previous section that after logarithmic number of steps, the anchor words will be correctly identified, and estimated within a factor of 2. We show that this is enough to cause the support of the discriminative words to be correctly identified too, as well as estimate them to within a constant factor where they are non-zero.

Same as before, we will assume in this section that we can identify the dominating topic.

We will crucially rely on the fact that the discriminative words will not have a very large dynamic range comparatively to their total probability mass in a topic. The high level outline will be similar to the case for the anchor words. We will prove that if a discriminative word j is in the support of topic i, then $\beta^t_{i,j}$ will always be reasonably lower bounded, and this will cause the values $\beta^t_{i',j}$ to keep decaying for the topics i' that the word j does not belong to.

The reason we will need the bound on the dynamic range, and the proportion of the dominating topic, and the size of the dominating topic, is to ensure that the β's are always properly lower bounded.

2.4.3.1 Bounds on the $\beta^t_{i,j}$ values

First, we show that because the discriminative words have a small range, the values $\beta^t_{i,j}$ whenever $\beta^*_{i,j}$ is non-zero are always maintained to be within some multiplicative constant (which depends on the range of the $\beta^*_{i,j}$).

As a preliminary, notice that having identified the anchor words correctly the γ values are appropriately lower bounded after running the γ update. Namely, by Lemma 35, $\gamma^t_{d,i} \geq p/2\gamma^*_{d,i}$

With this in hand, we show that the $\beta^t_{i,j}$ values are well upper bounded whenever $\beta^*_{i,j}$ is non-zero.

Lemma 36. *At any point in time t,* $\beta^t_{i,j} \leq (1+\epsilon)\frac{2B}{C_l}\beta^*_{i,j}$.

Proof. Since $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j}\gamma^t_{d,i} \leq \tilde{f}_{d,j}$ we have:

$$\beta^{t+1}_{i,j} \leq \frac{\sum_d \tilde{f}_{d,j}}{\sum_d \gamma^t_{d,i}} \leq 2 \cdot \frac{\sum_d \tilde{f}_{d,j}}{\sum_d \gamma^*_{d,i}}$$

On the other hand, we claim that $\tilde{f}_{d,j} \leq (1+\epsilon)B\beta^*_{i,j}$. Indeed, $\tilde{f}_{d,j} \leq (1+\epsilon)\sum_i \gamma^*_{d,i}\beta^*_{i,j}$, and for any other topic i', $\beta^*_{i',j} \leq B\beta^*_{i,j}$. Hence,

$$2 \cdot \frac{\sum_d \tilde{f}_{d,j}}{\sum_d \gamma^*_{d,i}} \leq \frac{2(1+\epsilon)DB\beta^*_{i,j}}{\sum_d \gamma^*_{d,i}}$$

However, since $\gamma^*_{d,i} \geq C_l$, the previous expression is at most

$$\frac{2(1+\epsilon)DB\beta^*_{i,j}}{DC_l} = \frac{2(1+\epsilon)B}{C_l}\beta^*_{i,j}$$

So, we get the claim we wanted.

□

The lower bound on the $\beta^t_{i,j}$ values is a bit more involved. To show a lower bound on the $\beta^t_{i,j}$ values is maintained, we will make use of both the fact that the discriminative words have a small range, and that we have some small, but reasonable proportion of documents where $\gamma^*_{d,i} \geq 1 - \delta$. More precisely, we show:

Lemma 37. *Let $\beta^t_{i,j} \leq \frac{2(1+\epsilon)B}{C_l}\beta^*_{i,j}$ for all topics i that word j belongs to, and let $\beta^t_{i,j} \geq \frac{C_l}{B}\beta^*_{i,j}$. Then, $\beta^{t+1}_{i,j} \geq \frac{C_l}{B}\beta^*_{i,j}$ as well.*

Proof. Let's call D_δ the documents where $\gamma^*_{d,i} \geq 1 - \delta$. We can certainly lower bound

$$\beta^{t+1}_{i,j} \geq \frac{\sum_{d \in D_\delta} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \gamma^t_{d,i} \beta^t_{i,j}}{\sum_{d \in D} \gamma^t_{d,i}}$$

First, let's focus on $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j}$. Then,

$$\tilde{f}_{d,j} \geq (1 - \epsilon)(1 - \delta)\beta^*_{i,j} \tag{2.4.3}$$

Furthermore, since $\sum_{d \in D_\delta} \gamma^t_{d,i} \geq \frac{1}{2} \sum_{d \in D_\delta} \gamma^*_{d,i}$ and $\sum_d \gamma^t_{d,i} \leq 2 \sum_d \gamma^*_{d,i}$, we have that

$$\frac{\sum_{d \in D_\delta} \gamma^t_{d,i}}{\sum_d \gamma^t_{d,i}} \geq \frac{1}{4}\frac{8}{B}(1 - \delta) = \frac{2}{B}(1 - \delta) \tag{2.4.4}$$

Finally, we claim that $\frac{\beta^t_{i,j}}{f^t_{d,j}} \geq \frac{1}{2}$. Massaging this inequality a bit, we get it's equivalent to:

$$\frac{\beta^t_{i,j}}{f^t_{d,j}} \geq \frac{1}{2} \Leftrightarrow$$

$$f^t_{i,j} \leq 2\beta^t_{i,j} \Leftrightarrow$$

$$\gamma^t_{d,i}\beta^t_{i,j} + \sum_{i'} \gamma^t_{d,i'}\beta^t_{i',j} \leq 2\beta^t_{i,j}$$

The left hand side can be upper bounded by

$$\gamma^t_{d,i}\beta^t_{i,j} + \sum_{i'} \gamma^t_{d,i'} \frac{2(1 + \epsilon)B^3}{C_l^2}\beta^t_{i,j} \leq$$

$$\gamma^t_{d,i}\beta^t_{i,j} + (1 - \gamma^t_{d,i})\frac{2(1 + \epsilon)B^3}{C_l^2}\beta^t_{i,j}$$

by the assumptions of the lemma.

So, it is sufficient to show that $\gamma^t_{d,i}\beta^t_{i,j} + (1 - \gamma^t_{d,i})\frac{2(1+\epsilon)B^3}{C_l^2}\beta^t_{i,j} \leq 2\beta^t_{i,j}$, however this is equivalent after some rearrangement to $\gamma^t_{d,i} \geq 1 - \frac{1}{\frac{2(1+\epsilon)B^3}{C_l^2} - 1}$.

It's certainly sufficient for this that $\gamma^t_{d,i} \geq 1 - \frac{1}{\frac{B^3}{C_l^2}} = 1 - \frac{C_l^2}{B^3}$, but since since $\gamma^*_{d,i} \geq 1 - \delta$, by the definition of δ and Lemmas 28, 39, 40, this certainly holds.

Together with 2.4.4 and 2.4.3, we get that

$$\beta^{t+1}_{i,j} \geq (1-\epsilon)\frac{2}{B}(1-\delta)^2\frac{1}{2}\beta^*_{i,j} \geq (1-\epsilon)\frac{(1-\delta)^2}{B}\beta^*_{i,j}$$

But, by our assumptions, $(1-\epsilon)(1-\delta)^2 \geq C_l$, so the claim follows.

□

2.4.3.2 Decreasing $\beta^t_{i',j}$ values

Finally, we show that if the discriminative word j does not belong in topic i', the value for $\beta^t_{i',j}$ will keep dropping. More precisely, the following is true:

Lemma 38. *Let word j and topic i be such that $\beta^*_{i',j} = 0$ and let $\beta^t_{i',j} \leq b$. Furthermore, let for all the topics i that j belongs to hold: $\beta^t_{i,j} \geq 1/C_\beta \beta^*_{i,j}$ for some constant C_β. Finally, let $\gamma^t_{d,i} \geq \frac{1}{C_\gamma}\gamma^*_{d,i}$ for some constant C_γ. Then, $\beta^{t+1}_{i',j} \leq b/2$.*

Proof. We proceed similarly as the analogous claim for anchor words. We split the update as

$$\beta^{t+1}_{i',j} = \beta^t_{i',j}\left(\frac{\sum_{d\in D_1}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}} + \frac{\sum_{d\in D_2}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i'}}{\sum_d \gamma^t_{d,i'}}\right)$$

for some appropriate partitioning of the documents D_1, D_2.

Namely, let D_1 be documents which do not contain any topic to which word j belongs, the D_2 documents which contain at least one topic word j belongs to.

For all the documents in D_1, $f^*_{d,j} = 0$, and we will provide a good bound for the terms $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}$ in D_2, this way, we'll ensure $\beta^t_{i,j}$ gets multiplied by a quantity which is $o(1)$ to get $\beta^{t+1}_{i,j}$, which is of course enough for what we want.

Bounding the terms in D_2 is even simpler than before. We have:

$$f^t_{d,j} = \sum_i \beta^t_{i,j}\gamma^t_{d,i} \geq \frac{1}{C_\beta C_\gamma}\sum_i \beta^*_{i,j}\gamma^*_{d,i} = \frac{1}{C_\beta C_\gamma} f^*_{d,j}$$

Hence, $\frac{f^*_{d,j}}{f^t_{d,j}} \leq C_\beta C_\gamma$.

Then we have:

$$\frac{\sum_d \frac{\tilde{f}_{d,j}}{f^t_{d,j}}\gamma^t_{d,i}}{\sum_d \gamma^t_{d,i}} \leq (1+\epsilon)\frac{\sum_d \frac{f^*_{d,j}}{f^t_{d,j}}\gamma^t_{d,i}}{\sum_d \gamma^t_{d,i}} \leq$$

$$4(1+\epsilon)\frac{\sum_d \frac{f^*_{d,j}}{f^t_{d,j}}\gamma^*_{d,i}}{\sum_d \gamma^*_{d,i}} \leq 4(1+\epsilon)\frac{\sum_{d\in D_2} C_\beta C_\gamma \gamma^*_{d,i}}{\sum_d \gamma^*_{d,i}}$$

But now, by the "weak topic correlation" property, $\frac{\sum_{d\in D_2}\gamma^*_{d,i}}{\sum_d \gamma^*_{d,i}} = o(1)$. Indeed, D consists of the documents where i' is the dominating topic. In order for the document to belong to D_2, at least one of the topics word j belongs to must belong in the document as well. Since the word j only belongs to $o(K)$ of the topics, and each document contains only a constant number of topics, by the small topic correlation property, the claim we want follows.

But then, clearly, $4\frac{\sum_{d\in D_2} C_\beta C_\gamma \gamma^*_{d,i}}{\sum_d \gamma^*_{d,i}} = o(1)$ as well.

Hence, $\beta^{t+1}_{i',j} = o(1)\beta^t_{i',j} \leq \frac{1}{2}\beta^t_{i',j}$, which is what we need. □

2.4.4 Determining dominant topic and parameter range

To complete the proofs of the claims for Phase I and II, we need to show that at any point in time we correctly identify the dominant topic. Furthermore, in order to maintain the lower bounds on the estimates for the discriminative words, we will need to make sure that $\gamma^t_{d,i}$ is large as well in the documents where $\gamma^*_{d,i} \geq 1-\delta$.

Let's proceed to the problem of detecting the largest topic first. By Lemma 29 all we need to do is bound R_f and R_β at any point in time during this phase. To do this, let's show the following lemma:

Lemma 39. *Suppose for the anchor words $\beta^t_{i,j} \geq C_1\beta^*_{i,j}$, for the discriminative words $\beta^t_{i,j} \geq C_2\beta^*_{i,j}$. Let p_i be the proportion of anchor words in topic i. Then, $KL(\beta^*_i||\beta^t_i) \leq p_i \log(\frac{1}{C_1}) + (1-p_i)\log(\frac{1}{C_2})$.*

Proof. This is quite simple. Since log is an increasing function,

$$KL(\beta^*_i||\beta^t_i) = \sum_j \beta^*_{i,j}\log(\frac{\beta^*_{i,j}}{\beta^t_{i,j}}) \leq p_i\log(\frac{1}{C_1}) + (1-p_i)\log(\frac{1}{C_2})$$

□

Lemma 40. *Suppose for the anchor words $\beta^t_{i,j} \geq C_1\beta^*_{i,j}$, for the discriminative words $\beta^t_{i,j} \geq C_2\beta^*_{i,j}$. Let p_i be the proportion of anchor words in topic i. Then, $\min_{\gamma\in\Delta_K} KL(\tilde{f}_d||f_d) \leq \log(1+\epsilon) + \left(p\log(\frac{1}{C_1}) + (1-p)\log(\frac{1}{C_2})\right)$.*

Proof. Also simple. The value of $KL(\tilde{f}_d||f_d)$ one gets by plugging in $\gamma_d = \gamma^*$ is exactly what is stated in the lemma.

□

We'll just use the above two lemmas combined from our estimates from before. We know, for all the anchor words, that $\beta^t_{i,j} \geq C_l\beta^*_{i,j}$, and that for the discriminative words, $\beta^t_{i,j} \geq \frac{C_l}{B}\beta^*_{i,j}$. Hence, by Lemma 39, at any point in time $KL(\beta^*_i||\beta^t_i) \leq p\log(\frac{1}{C_l}) + (1-p)\log(\frac{B}{C_l})$. So, by Lemma 29, it's enough that

$$C_l - C_s \geq \frac{1}{p}\left(\sqrt{2\left(p\log(\frac{1}{C_l}) + (1-p)\log(BC_l)\right)} + \sqrt{\log(1+\epsilon)}\right) + \epsilon$$

.

Since $\frac{1}{p}\sqrt{2\left(p\log(\frac{1}{C_l})+(1-p)\log(BC_l)\right)} \le \frac{1}{p}\sqrt{2\left(\log(\frac{1}{C_l})+(1-p)\log B\right)}$, to get a sense of the parameters one can achieve, for detecting the dominant topic, (ignoring ϵ contributions), it's sufficient that $C_l - C_s \ge \frac{2}{p}\sqrt{\max(\log(\frac{1}{C_l}),(1-p)\log B)}$

If one thinks of C_l as $1-\eta$ and $p \ge 1-\frac{\eta}{\log B}$, since $\log(\frac{1}{C_l}) \approx \eta$ roughly we want that $C_l - C_s \gg \frac{2}{p}\sqrt{\eta}$. (One takeaway message here is that the weight we require to have on the anchors depends only *logarithmically* on the range B.)

Let's finally figure out what the topic proportions must be in the "heavy" documents. In these, we want $\gamma^*_{d,i} \ge 1 - \frac{C_l^2}{2B^3} + \frac{1}{p}\left(\sqrt{2\left(p\log(\frac{1}{C_l})+(1-p)\log(BC_l)\right)} - \sqrt{\log(1+\epsilon)}\right) + \epsilon$. A similar approximation to the above gives that we roughly want $\gamma^*_{d,i} \ge 1 - \frac{1-2\eta}{2B^3} + \frac{2}{p}\sqrt{\eta}$.

2.4.5 Getting the supports correct

At the end of the previous section, we argued that after $O(\log n)$ rounds, we will identify the anchor words correctly, and the supports of the discriminative words as well. Furthermore, we will also have estimated the values of the non-zero discriminative word probabilities, as well the anchor word probabilities up to a multiplicative constant. Then, I claim that from this point onward at each of the γ steps, the γ^t values we get will have the correct support. Namely, the following is true:

Lemma 41. *Suppose for the anchor words and discriminative words j, if $\beta^*_{i,j} = 0$, it's true that $\beta^t_{i,j} = o(\frac{1}{n})$. Furthermore, suppose that if $\beta^*_{i,j} \neq 0$, $\frac{1}{C_\beta}\beta^*_{i,j} \le \beta^t_{i,j} \le C_\beta\beta^*_{i,j}$ for some constant C_β.*

Then, when performing KL minimization with respect to the γ variables, whenever $\gamma^*_{d,i} = 0$ we have $\gamma^t_{d,i} = 0$.

Proof. Let $\gamma^*_{d,i} = 0$. If $\gamma^t_{d,i} \neq 0$, then the KKT conditions imply:

$$\sum_{j=1}^{n} \frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} = 1 \tag{2.4.5}$$

The only terms that are non-zero in the above summation are due to words j that belong to at least one topic i' in the document. Let I be the set of words that belong to topic i as well.

By Lemma 35, we know that $\gamma^t_{d,i} \ge p/2\gamma^*_{d,i}$ Since also $\beta^t_{i,j} \ge \frac{1}{C_\beta}\beta^*_{i,j}$, $f^t_{d,i} \ge \frac{p}{2C_\beta}f^*_{d,j}$. Since $\beta^t_{i,j} = o(\frac{1}{n})$ for words j not in the support of topic I, $\sum_{j\notin I}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} = o(1)$.

On the other hand, for words in I, $\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} \le (1+\epsilon)\frac{2C_\beta^2}{p}\beta^*_{i,j}$, so $\sum_{j\in I}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} = o(1)$, by the small support intersection property.

However, this contradicts 2.4.5, so we get what we want.

□

This means that after this phase, we will always correctly identify the supports of the γ variables as well.

2.4.6 Alternating minimization

Now, finishing the proof of Theorem 24 is trivial. Namely, because of Lemmas 41, 33, and the analogue of 33, we are basically back to the case where we have the correct supports for both the β and γ variables. The only thing left to deal with is the fact that the β variables are not quite zero.

Let j be an anchor word for topic i. Let $\epsilon'' = 1 - (1 - \epsilon')^{1/7}$. Similarly as in Lemma 35, for

$$t > 10 \max(\log n, \log(\frac{1}{\epsilon'' \gamma^*_{\min}}), \log(\frac{1}{\epsilon'' \beta^*_{\min}}))$$

it holds that $\frac{f^*_{d,j}}{f^t_{d,j}} \geq (1 - \epsilon')^{1/7} \frac{\beta^*_{i,j}\gamma^*_{d,i}}{\beta^t_{d,i}\gamma^t_{d,i}}$. The same inequality is true if j is a lone word for topic i in document d.

After the above event, the same proof from Case Study 1 implies that after $O(\log(\frac{1}{\epsilon'}))$ iterations we'll get

$$\frac{1}{1+\epsilon'} \beta^*_{i,j} \leq \beta^t_{i,j} \leq (1 + \epsilon')\beta^*_{i,j}$$

and

$$\frac{1}{1+\epsilon'} \gamma^*_{i,j} \leq \gamma^t_{i,j} \leq (1 + \epsilon')\gamma^*_{i,j}$$

This finishes the proof of Theorem 24.

2.5 On common words

We finally turn to a few extensions involving *common words*: words such that $\beta^*_{i,j} \leq \kappa \beta^*_{i',j}, \forall i, i', \kappa \leq B$.

In this section, we show how one would modify the proofs from the previous section to handle common words as well. We stress that common words are easy to handle if one were allowed to filter them out (and this in fact, is often done in practice), but we want to analyze under which conditions the variational inference updates could handle them on their own. The difference in contrast to the previous sections is it's not clear how to argue progress for the common words: common words do not have lone documents. However, if we can't argue progress for the common words, then we can't argue progress for the γ variables, so the entire argument seems to fail.

Concretely, we consider the following scenario:

- On top of the assumptions we have either in Case Study 1 or Case Study 2, we assume that there are words which show up in all topics, but their probabilities are within a constant κ from each other, $B \geq \kappa \geq 2$. We will call these *common* words. (The $\kappa \geq 2$ is without loss of generality. If the claim holds for a smaller κ, then it certainly holds for $\kappa = 2$. The only difference is that the estimates to follow could be strengthened, but we assume $\kappa \geq 2$ to get cleaner bounds.)

- For each topic i, if C is the set of common words, $\sum_{j \in C} \beta^*_{i,j} \leq \frac{1}{\kappa^{100}}$, i.e. there isn't too much mass on these words.
- Conditioned on topic i being dominant, there is a probability of $1 - \frac{1}{\kappa^{100}}$ that the proportion of topic i is at least $1 - \frac{1}{\kappa^{100}}$.

Then, the theorem we can prove is:

Theorem 42. *If we additionally have common words satisfying the properties specified above, after $O(\log(1/\epsilon') + \log N)$ of KL-tEM updates in Case Study 2, or any of the tEM variants in Case Study 1, and we use the same initializations as before, we recover the topic-word matrix and topic proportions to multiplicative accuracy $1 + \epsilon'$, if $1 + \epsilon' \geq \frac{1}{(1-\epsilon)^7}$.*

Our bounds and analysis here is fairly loose, since the result is anyway somewhat weak. (e.g. $1 - \frac{1}{\kappa^{100}}$ is not really the best value for the proportion of the dominating topic, or the proportion of such documents required.) At any rate, it will be clear from the proofs that the dependency of the dominating topic on κ has to be of the form $1 - \frac{1}{\kappa^c}$, so it's not clear one would gain too much from the tightest possible analysis. The reason we are including this section is to show cases where these proof methods start breaking down.

We will do the proof for Case Study 1 first, after which Case Study 2 will easily follow.

2.5.1 Phase I with common words

The outline is the same as before. We prove the lower bounds on the γ and β variables first. Namely, we prove:

Lemma 43. *Suppose that the supports of β and γ are correct. Then, $\gamma^t_{d,i} \geq \frac{1}{2}\gamma^*_{d,i}$.*

Proof. Similarly as before, multiplying both sides of 2.3.1 by $\gamma^t_{d,i}$, we get that

$$\gamma^t_{d,i} \geq \sum_{L_i} \frac{f^*_{d,j}}{f^t_{d,j}} \beta^t_{i,j} \gamma^t_{d,i} \geq (1 - o(1))(1 - \frac{1}{\kappa^{100}})\gamma^*_{d,i} \geq \frac{1}{2}\gamma^*_{d,i}$$

where the second inequality follows since $1 - \frac{1}{\kappa^{100}}$ fraction of the words in topic i is discriminative. □

Lemma 44. *Suppose that the supports of the γ and β variables are correct. Additionally, if i is a large topic in d, let $\frac{1}{2}\gamma^*_{d,i} \leq \gamma^t_{d,i} \leq 3\gamma^*_{d,i}$. Then, for a discriminative word j for topic i, $\beta^{t+1}_{i,j} \geq \frac{1}{3}\beta^*_{i,j}$.*

Proof. Again, similarly as in Lemma 10,

$$\beta^{t+1}_{i,j} \geq \frac{\sum_{d \in D_l} (1 - \epsilon) \frac{\gamma^*_{d,i}\beta^*_{i,j}}{\gamma^t_{d,i}\cdot\beta^t_{i,j}} \beta^t_{i,j} \gamma^t_{d,i}}{\sum_{d=1}^{D} \gamma^t_{d,i}} =$$

$$(1-\epsilon)\beta^*_{i,j}\frac{\sum_{d\in D_l}^{D}\gamma^*_{d,i}}{\sum_{d=1}^{D}\gamma^t_{d,i}}$$

In the documents where topic i is the largest, $\gamma^t_{d,i} \le 3\gamma^*_{d,i}$. So, we can conclude

$$\beta^{t+1}_{i,j} \ge (1-\epsilon)\beta^*_{i,j}\frac{1}{3}\frac{\sum_{d\in D_l}^{D}\gamma^*_{d,i}}{\sum_{d=1}^{D}\gamma^*_{d,i}}$$

Since $\frac{\sum_{d\in D_l}^{D}\gamma^*_{d,i}}{\sum_{d=1}^{D}\gamma^*_{d,i}} \ge (1-o(1))$, as before, we get what we want.

□

Lemma 45. *Let the β variables have the correct support. Let j be a discriminative word for topic i, and let $\beta^t_{i,j} \ge \frac{1}{C_m}\beta^*_{i,j}$, $\gamma^t_{d,i} \ge \frac{1}{C_m}\gamma^*_{d,i}$ whenever $\beta^*_{i,j} \ne 0$, $\gamma^*_{d,i} \ne 0$. Let $\beta^t_{i,j} = C^t_\beta\beta^*_{i,j}$, where $C^t_\beta \ge 4C_m$, and C_m is a constant. Then, in the next iteration, $\beta^{t+1}_{i,j} \le C^{t+1}_\beta\beta^*_{i,j}$, where $C^{t+1}_\beta \le \frac{C^t_\beta}{2}$.*

Proof. The proof is exactly the same as Lemma 11.

□

Now, we finally get to the upper bound of the γ values.

Lemma 46. *Fix a particular document d. Let's assume the supports for the β and γ variables are correct. Furthermore, let $\frac{1}{C_m} \le \frac{\beta^t_{i,j}}{\beta^*_{i,j}} \le C_m$ for some constant C_m. Then, $\gamma^t_{d,i} \le 2\gamma^*_{d,i}$.*

Proof. Again, multiplying 2.3.1 by $\gamma^t_{d,i}$, we get

$$\gamma^t_{d,i} = \sum_{j\in L_i}\tilde{f}_{d,j} + \gamma^t_{d,i}\sum_{j\notin L_i}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j} + \gamma^t_{d,i}\sum_{j\in C}\frac{\tilde{f}_{d,j}}{f^t_{d,j}}\beta^t_{i,j}$$

If $\tilde{\alpha} = \sum_{j\in L_i}\beta^*_{i,j}$, since $\gamma^t_{d,i} \ge \frac{1}{C_m}\gamma^*_{d,i}$,

$$\frac{\tilde{f}_{d,j}}{f^t_{d,j}} \le (1+\epsilon)C_m^2$$

If we denote $\Gamma = \sum_{j\in C}\beta^*_{i,j}$, then

$$\gamma^t_{d,i} \le (1+\epsilon)(\tilde{\alpha}\gamma^*_{d,i} + C_m^3(1-\Gamma-\tilde{\alpha})\gamma^t_{d,i} + \Gamma\kappa^4\gamma^t_{d,i})$$

Equivalently, $\gamma^t_{d,i} \le \frac{(1+\epsilon)\tilde{\alpha}}{1-(1+\epsilon)C_m^3(1-\Gamma-\tilde{\alpha})-(1+\epsilon)\Gamma\kappa^4}\gamma^*_{d,i}$

Then, we claim that $\frac{(1+\epsilon)\tilde{\alpha}}{1-(1+\epsilon)C_m^3(1-\Gamma-\tilde{\alpha})-(1+\epsilon)\Gamma\kappa^4} \le 1+\frac{1}{\kappa^{50}}$. Indeed, $\Gamma\kappa^4 \le \kappa^{-96}$, and $C_m^3(1-\Gamma-\tilde{\alpha}) \le C_m^3(1-\tilde{\alpha}) = o(1)$. Hence,

$$\frac{(1+\epsilon)\tilde{\alpha}}{1-(1+\epsilon)C_m^3(1-\Gamma-\tilde{\alpha})-(1+\epsilon)\Gamma\kappa^4} \le \frac{(1+\epsilon)\tilde{\alpha}}{1-o(1)-\kappa^{-96}} \le \frac{(1+\epsilon)\tilde{\alpha}}{1-\kappa^{-95}}$$

Finally, we claim that $\frac{(1+\epsilon)\tilde{\alpha}}{1-\kappa^{-95}} \leq 1+\kappa^{-50}$. Indeed, this is equivalent to

$$\tilde{\alpha} \leq (1+\epsilon)(1+\kappa^{-50})(1-\kappa^{-95}) \leq (1+\epsilon)(1+\kappa^{-50})$$

But, since we assume $\kappa \geq 2$, the claim we need follows easily. □

2.5.2 Phase II of analysis

Finally, we deal with the alternating minimization portion of the argument. How will we deal with the lack of anchor documents? The almost obvious way: if a document has topic i with proportion $1-\frac{1}{\kappa^{100}}$, it will behave for all purposes like an anchor document, because the dynamic range of word $\beta^*_{i,j}$ is limited, and the contribution from the other topics is not that significant.

Intuitively, we'll show that $\frac{f^*_{d,j}}{f^t_{d,j}} \approx \frac{\beta^*_{i,j}}{\beta^t_{i,j}}$, so that these documents provide a "push" for the value of $\beta^t_{i,j}$ in the correct direction.

Lemma 47. *Let's assume that our current iterates $\beta^t_{i,j}$ satisfy $\frac{1}{C^t_\beta} \leq \frac{\beta_{i,j}*}{\beta^t_{i,j}} \leq C^t_\beta$ for $C^t_\beta \geq \frac{1}{(1-\epsilon)^{20}}$. Then, after iterating the γ updates to convergence, we will get values $\gamma^t_{d,i}$ that satisfy $(C^t_\beta)^{1/10} \leq \frac{\gamma_{d,i}*}{\gamma^t_{d,i}} \leq (C^t_\beta)^{1/10}$.*

Proof. As before, we have that

$$\gamma^t_{d,i} = \sum_{j \in L_i} \tilde{f}_{d,j} + \gamma^t_{d,i} \sum_{j \notin L_i} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \beta^t_{i,j}$$

Let's denote as $C^t_\gamma = \max_i(\max(\frac{\gamma^*_{d,i}}{\gamma^t_{d,i}}, \frac{\gamma^t_{d,i}}{\gamma^*_{d,i}}))$, and let, as before, assume that $\frac{\gamma^t_{d,i_0}}{\gamma^*_{d,i_0}} = C^t_\gamma$.

By the definition of C^t_γ,

$$\gamma^t_{d,i_0} = \sum_{j \in L_{i_0}} \tilde{f}_{d,j} + \gamma^t_{d,i_0} \sum_{j \notin L_{i_0}} \frac{\tilde{f}_{d,j}}{f^t_{d,j}} \beta^t_{i_0,j} \leq$$

$$(1+\epsilon)(\tilde{\alpha}\gamma^*_{d,i_0} + (1-\tilde{\alpha})(C^t_\beta)^2 (C^t_\gamma)^2 \gamma^*_{d,i_0})$$

We claim that

$$(1+\epsilon)(\tilde{\alpha} + (1-\tilde{\alpha})(C^t_\beta)^2 (C^t_\gamma)^2) \leq (C^t_\gamma)^{1/10} \tag{2.5.1}$$

which will be a contradiction to the definition of C^t_γ.

After a little rewriting, 2.5.1 translates to $\tilde{\alpha} \geq 1 - \frac{\frac{(C^t_\gamma)^{1/10}}{1+\epsilon}-1}{(C^t_\beta C^t_\gamma)^2-1}$. By our assumption on C^t_γ, $C^t_\beta \leq C^{10}_\gamma$, so the right hand side above is upper bounded by $1 - \frac{\frac{(C^t_\gamma)^{1/10}}{1+\epsilon}-1}{(C^t_\gamma)^8-1}$.

But, Lemma 46 implies that certainly $C_\gamma^t \le C_\gamma^0$. The function

$$f(c) = \frac{\frac{c^{1/10}}{1+\epsilon} - 1}{c^8 - 1}$$

can be easily seen to be monotonically decreasing on the interval of interest, and hence is lower bounded by $\frac{\frac{(C_\gamma^0)^{1/10}}{1+\epsilon}-1}{(C_\gamma^0)^8-1}$. Since $\tilde{\alpha} = (1 - o(1))(1 - \frac{1}{\kappa^{100}})$ and $C_\gamma^0 \le 3$, the claim we want is clearly true.

The case where $\frac{\gamma^*_{d,i_0}}{\gamma^t_{d,i_0}} = C_\gamma^t$ is not much more difficult. An analogous calculation as in Lemma 14 gives that to get a contradiction to the definition of C_γ^t, the condition required is that $1 - \frac{1-\frac{1}{(1-\epsilon)(C_\gamma^t)^{1/10}}}{1-\frac{1}{(C_\gamma^t)^8}}$. As before, if $f(c) = \frac{1-\frac{1}{(1-\epsilon)c^{1/10}}}{1-c^8}$, it s easy to check that $f(c)$ is monotonically increasing in the interval of interest, so lower bounded by

$$\frac{1 - \frac{1}{(1-\epsilon)(\frac{1}{(1-\epsilon)^{20}})^{1/10}}}{1 - \frac{1}{((\frac{1}{1-\epsilon})^{20})^8}} =$$

$$\frac{1 - (1-\epsilon)}{1 - (1-\epsilon)^{160}} \ge \frac{1}{160}$$

But, $\tilde{\alpha} \ge (1 - \frac{1}{\kappa^{100}})(1 - o(1)) \ge 1 - \frac{1}{160}$, so we get what we want.

□

Next, we show the following lemma.

Lemma 48. *Suppose at time step t, $\frac{1}{C_\gamma^t}\gamma^*_{d,i} \le \gamma^t_{d,i} \le C_\gamma^t \gamma^*_{d,i}$ and $\frac{1}{C_\beta^t}\beta^*_{i,j} \le \beta^t_{i,j} \le C_\beta^t \beta^*_{i,j}$, such that $C_\gamma^t \le (C_\beta^t)^{1/10}$ for $C_\beta^t \ge \frac{1}{(1-\epsilon)^{20}}$. Then, at time step $t+1$, $1/C_\beta^{t+1}\beta^*_{i,j} \le \beta^t_{i,j} \le C_\beta^{t+1}\beta^*_{i,j}$, where $C_\beta^{t+1} = (C_\beta^t)^{3/4}$*

Proof. Let's assume a document d has a dominating topic of proportion at least $1 - 1/\kappa^{100}$.

Then, we claim that $\frac{f^*_{d,j}}{f^t_{d,j}} \ge \frac{1}{(C_\beta^t)^{1/4}}\frac{\beta^*_{i,j}}{\beta^t_{i,j}}$. We will do a sequence of rearrangements to get this condition to a simpler form:

$$\frac{f^*_{d,j}}{f^t_{d,j}} \ge \frac{1}{(C_\beta^t)^{1/4}}\frac{\beta^*_{i,j}}{\beta^t_{i,j}} \Leftrightarrow$$

$$\frac{f^*_{d,j}}{\beta^*_{i,j}} \ge \frac{1}{(C_\beta^t)^{1/4}}\frac{f^t_{d,j}}{\beta^t_{i,j}} \Leftrightarrow$$

$$\gamma^*_{d,i} + \sum_{i'} \gamma^*_{d,i'}\frac{\beta^*_{i',j}}{\beta^*_{i,j}} > \frac{1}{(C_\beta^t)^{1/4}}(\gamma^t_{d,i} + \sum_{i'} \gamma^t_{d,i'}\frac{\beta^t_{i',j}}{\beta^t_{i,j}})$$

Let's upper bound the right hand side by some simpler quantities. We have:

$$\frac{1}{(C_\beta^t)^{1/4}}(\gamma^t_{d,i} + \sum_{i'} \gamma^t_{d,i'}\frac{\beta^t_{i',j}}{\beta^t_{i,j}}) \le$$

$$\frac{1}{(C_\beta^t)^{1/4}} C_\gamma^t (\gamma_{d,i}^* + \sum_{i'} \gamma_{d,i'}^* \frac{\beta_{i',j}^t}{\beta_{i,j}^t}) \leq$$

$$\frac{1}{(C_\beta^t)^{1/4}} C_\gamma^t (\gamma_{d,i}^* + (C_\beta^t)^2 \sum_{i'} \gamma_{d,i'}^* \frac{\beta_{i',j}^*}{\beta_{i,j}^*})$$

Hence, it is sufficient to prove

$$\gamma_{d,i}^* + \sum_{i'} \gamma_{d,i'}^* \frac{\beta_{i',j}^*}{\beta_{i,j}^*} \geq \frac{1}{(C_\beta^t)^{1/4}} C_\gamma^t (\gamma_{d,i}^* + (C_\beta^t)^2 \sum_{i'} \gamma_{d,i'}^* \frac{\beta_{i',j}^*}{\beta_{i,j}^*}) \Leftrightarrow$$

$$\gamma_{d,i}^* (1 - \frac{C_\gamma^t}{(C_\beta^t)^{1/4}}) \geq \sum_{i'} \gamma_{d,i'}^* (\frac{C_\gamma^t}{(C_\beta^t)^{1/4}} (C_\beta^t)^2 - 1) \frac{\beta_{i',j}^*}{\beta_{i,j}^*}$$

Again, we can upper bound the right hand side by

$$\sum_{i'} \gamma_{d,i'}^* (\frac{C_\gamma^t}{(C_\beta^t)^{1/4}} (C_\beta^t)^2 - 1)\kappa =$$

$$(1 - \gamma_{d,i}^*)(\frac{C_\gamma^t}{(C_\beta^t)^{1/4}} (C_\beta^t)^2 - 1)\kappa$$

So, it is sufficient to prove:

$$(1 - \gamma_{d,i}^*)(\frac{C_\gamma^t}{(C_\beta^t)^{1/4}} (C_\beta^t)^2 - 1)\kappa \leq \gamma_{d,i}^* (1 - \frac{C_\gamma^t}{(C_\beta^t)^{1/4}}) \Leftrightarrow$$

$$\gamma_{d,i}^* (1 - \frac{C_\gamma^t}{(C_\beta^t)^{1/4}} + (\frac{C_\gamma^t}{(C_\beta^t)^{1/4}} (C_\beta^t)^2 - 1)\kappa) \geq (\frac{C_\gamma^t}{(C_\beta^t)^{1/4}} (C_\beta^t)^2 - 1)\kappa \Leftrightarrow$$

$$\gamma_{d,i}^* \geq 1 - \frac{1 - \frac{C_\gamma^t}{(C_\beta^t)^{1/4}}}{1 - \frac{C_\gamma^t}{(C_\beta^t)^{1/4}} + (\frac{C_\gamma^t}{(C_\beta^t)^{1/4}} (C_\beta^t)^2 - 1)\kappa}$$

It's easy to check that the expression on the right hand side as a function of C_γ^t is decreasing. Hence, the RHS is upper bounded by

$$1 - \frac{1 - \frac{1}{(C_\beta^t)^{3/20}}}{1 - \frac{1}{(C_\beta^t)^{3/20}} + \kappa((C_\beta^t)^{37/20} - 1)}$$

Now, let's analyze this expression. If we let $f(x) = 1 - \frac{1 - \frac{1}{x^{3/20}}}{1 - \frac{1}{x^{3/20}} + \kappa(x^{37/20} - 1)}$, I claim $f(x)$ is an increasing function of x. Indeed, we can calculate it's derivative fairly easily:

$$f'(x) = -\frac{\frac{3}{20} x^{-\frac{23}{20}} (1 - \frac{1}{x^{3/20}} + \kappa(x^{37/20} - 1)) - (1 - \frac{1}{x^{3/20}})(-\frac{3}{20} x^{-\frac{23}{20}} + \frac{37}{20} \kappa x^{\frac{17}{20}})}{(1 - \frac{1}{x^{3/20}} + \kappa(x^{37/20} - 1))^2} =$$

$$-\frac{\frac{3}{20}x^{-\frac{23}{20}}\kappa(x^{\frac{37}{20}}-1)-\frac{37}{20}\kappa x^{\frac{17}{20}}(1-x^{-\frac{3}{20}})}{(1-\frac{1}{x^{3/20}}+\kappa(x^{37/20}-1))^2}=\frac{\frac{\kappa}{20}(40x^{14/20}-(3x^{-23/40}+37x^{17/20}))}{(1-x^{3/20}+\frac{1}{\kappa}(\frac{1}{x^{37/20}}-1))^2}$$

By the AM-GM inequality, $3x^{-23/40}+37x^{17/20}\geq 40((x^{17/20})^3 7(x^{-23/20})^3)^{1/40}=40x^{14/20}$, so $f'(x)$ is positive, so the RHS, as a function of C_β^t, is (x) is increasing.

So, it is sufficient to satisfy the inequality when $C_\beta^t=C_\beta^0$. One can check however that by Lemma 44 and 45 this is true.

Proceeding to the lower bound, a similar calculation as before gives that the necessary condition for progress is:

$$\gamma_{d,i}^*\geq 1-\frac{1-\frac{(C_\beta^t)^{1/4}}{C_\gamma^t}}{1-\frac{(C_\beta)^{1/4}}{C_\gamma^t}+\frac{1}{\kappa}(\frac{(C_\beta^t)^{1/4}}{C_\gamma^t}\frac{1}{(C_\beta^t)^2}-1)}$$

Again, the right hand side expression is decreasing in C_γ, so it is certainly upper bounded by

$$1-\frac{1-(C_\beta^t)^{3/20}}{1-(C_\beta^t)^{3/20}+\frac{1}{\kappa}(\frac{1}{(C_\beta^t)^{37/20}}-1)}$$

Now, the claim is that this expression is increasing in C_β^t. Again, denoting $f(x)=1-\frac{1-x^{3/20}}{1-x^{3/20}+\frac{1}{\kappa}(\frac{1}{x^{37/20}}-1)}$

$$f'(x)=-\frac{-\frac{3}{20}x^{-17/20}(1-x^{3/20}+\frac{1}{\kappa}(\frac{1}{x^{37/20}}-1))-(1-x^{3/20})(-\frac{3}{20}x^{-17/20}-\frac{1}{\kappa}\frac{37}{20}x^{-57/20})}{(1-x^{3/20}+\frac{1}{\kappa}(\frac{1}{x^{37/20}}-1))^2}=$$

$$-\frac{-\frac{3}{20}x^{-17/20}\frac{1}{\kappa}(\frac{1}{x^{37/20}}-1)+(1-x^{3/20})\frac{1}{\kappa}\frac{37}{20}x^{-57/20}}{(1-x^{3/20}+\frac{1}{\kappa}(\frac{1}{x^{37/20}}-1))^2}=\frac{\frac{1}{20\kappa}(-40x^{-54/20}+(3x^{-17/40}+37x^{-57/20}))}{(1-x^{3/20}+\frac{1}{\kappa}(\frac{1}{x^{37/20}}-1))^2}$$

By the AM-GM inequality, $3x^{-17/40}+37x^{-57/20}\geq 40((x^{-17/20})^3 7(x^{-57/20})^{37})^{1/40}=40x^{-54/20}$, so $f'(x)$ is negative, so the RHS, as a function of C_β^t, is decreasing. So it suffices to check the inequality when $C_\beta^t=(1-\epsilon)^{20}$. In this case, we want to check that

$$1-\frac{1}{\kappa^{100}}\geq 1-\frac{1-\frac{1}{(1-\epsilon)^3}}{1-\frac{1}{(1-\epsilon)^3}+\frac{1}{\kappa}((1-\epsilon)^{37}-1)}$$

Since $1-\frac{1-\frac{1}{(1-\epsilon)^3}}{1-\frac{1}{(1-\epsilon)^3}+\frac{1}{\kappa}((1-\epsilon)^{37}-1)}\leq 1-\frac{3\kappa}{37+3\kappa}$, and $\kappa\geq 2$, this is easily seen to be true.

Now, we'll split the β update into two parts: documents where topic i is at least $1-1/\kappa^{100}$, and the rest of them. In the first group, as we showed above, $\frac{f_{d,j}^*}{f_{d,j}^t}\geq\frac{1}{(C_\beta^t)^{1/2}}$. In the second group, we can certainly claim that $\frac{f_{d,j}^*}{f_{d,j}^t}\geq\frac{1}{C_\gamma^t C_\beta^t}$ from the inductive hypothesis. If we denote the set of documents where topic i is at least $1-1/\kappa^{100}$ as D_1, we get that

$$\beta_{i,j}^{t+1}=\beta_{i,j}^t\frac{\sum_d\frac{f_{d,j}^*}{f_{d,j}^t}\gamma_{d,i}^t}{\sum_{i=1}^D\gamma_{d,i}^t}\geq$$

$$\frac{\sum_{d\in D_1}\frac{1}{(C_\beta^t)^{1/2}C_\gamma^t}\beta_{i,j}^*\gamma_{d,i}^*+\sum_{d\in D\setminus D_1}\frac{1}{(C_\beta^t)^2(C_\gamma^t)^2}\beta_{i,j}^*\gamma_{d,i}^*}{(C_\gamma^t)\sum_{d\in D}\gamma_{d,i}^*}$$

If we denote $\mu = \frac{\sum_{d\in D_1}\gamma^*_{d,i}}{\sum_{d\in D}\gamma^*_{d,i}}$, then

$$\beta^{t+1}_{i,j} \geq \mu\frac{\beta^*_{i,j}}{(C^t_\beta)^{1/4}(C^t_\gamma)^2} + (1-\mu)\frac{\beta^*_{i,j}}{(C^t_\beta)^2(C^t_\gamma)^3}$$

So, to prove $\beta^{t+1}_{i,j} \geq \frac{1}{C^{3/4}_\beta}\beta^*_{i,j}$, it's sufficient to show

$$\mu\frac{\beta^*_{i,j}}{(C^t_\beta)^{1/4}(C^t_\gamma)^2} + (1-\mu)\frac{\beta^*_{i,j}}{(C^t_\beta)^2(C^t_\gamma)^3} \geq \frac{1}{C^{3/4}_\beta} \Leftrightarrow$$

$$\mu > \frac{\frac{1}{(C^t_\beta)^{1/2}} - \frac{1}{(C^t_\beta)^2(C^t_\gamma)^3}}{\frac{1}{(C^t_\beta)^{1/4}(C^t_\gamma)^2} - \frac{1}{(C^t_\beta)^2(C^t_\gamma)^3}}$$

Given that $C^t_\gamma \leq (C^t_\beta)^{1/10}$, it's sufficient to show

$$\mu > \frac{\frac{1}{(C^t_\beta)^{1/2}} - \frac{1}{(C^t_\beta)^{23/10}}}{\frac{1}{(C^t_\beta)^{9/20}} - \frac{1}{(C^t_\beta)^{23/10}}} = 1 - \frac{\frac{1}{(C^t_\beta)^{9/20}} - \frac{1}{(C^t_\beta)^{1/2}}}{\frac{1}{(C^t_\beta)^{9/20}} - \frac{1}{(C^t_\beta)^{23/10}}}$$

Completely analogously as before, $1 - \frac{\frac{1}{(C^t_\beta)^{9/20}} - \frac{1}{(C^t_\beta)^{1/2}}}{\frac{1}{(C^t_\beta)^{9/20}} - \frac{1}{(C^t_\beta)^{23/10}}}$ is a decreasing function of C^t_β, so it's sufficient to check that $\mu > 1 - \frac{\frac{1}{(C^t_\beta)^{9/20}} - \frac{1}{(C^t_\beta)^{1/2}}}{\frac{1}{(C^t_\beta)^{9/20}} - \frac{1}{(C^t_\beta)^{23/10}}}$ when $C^t_\beta = (\frac{1}{1-\epsilon})^{20}$, which is easily checked to be true.

In the same way, one can prove that $\beta^{t+1}_{i,j} \leq (C^t_\beta)^{3/4}\beta^*_{i,j}$ □

Putting lemmas 47 and 48 together, we get that the analogue of Lemma 16:

Lemma 49. *Suppose it holds that* $\frac{1}{C^t} \leq \frac{\beta_{i,j}*}{\beta^t_{i,j}} \leq C^t$, $C^t \geq \frac{1}{(1-\epsilon)^{20}}$. *Then, after one KL minimization step with respect to the* γ *variables and one* β *iteration, we get new values* $\beta^{t+1}_{i,j}$ *that satisfy* $\frac{1}{C^{t+1}} \leq \frac{\beta_{i,j}*}{\beta^{t+1}_{i,j}} \leq C^{t+1}$, *where* $C^{t+1} = (C^t)^{3/4}$

As a corollary,

Corollary 50. *Phase III requires* $O(\log(\frac{1}{\log(1+\epsilon)})) = O(\log(\frac{1}{\epsilon}))$ *iterations to estimate each of the topic-word matrix and document proportion entries to within a multiplicative factor of* $\frac{1}{(1-\epsilon)^7}$

This finished the proof of Theorem 42 for Case Study 1.

2.5.3 Generalizing Case Study 2

Finally, the proof for Case Study 2 is quite simple. Because the dynamic range $\kappa \leq B$ for the common words, Lemmas 39 and 40 still hold, and hence we again determine the dominant topic correctly. Because of this, it's also easy to see that the lower bounds and upper bounds on the $\beta^t_{i,j}$ values for the common words are maintained to be a

constant, since the proof of Lemmas 36 and 37 holds for the common words verbatim. This means that the anchor words and discriminative words will be correctly determined just as before. But after that point, the analysis of Case Study 2 is exactly the same as the one for Case Study 2 — which we already covered in the above section. This finishes the proof of Theorem 42.

2.6 Justification of prior assumptions via analogy to Dirichlet priors

In this section we provide a brief motivation for our choice of properties on the topic model instances we are looking at. Nothing in the other sections crucially depends on this section, so it can be freely skipped upon first reading.

Most of our properties on the topic priors are inspired from what happens with the Dirichlet prior - specifically, variants of all of the "weak correlations" between topics hold for Dirichlet. Essentially the only difference between our assumptions and Dirichlet is the lack of smoothness. (Dirichlet is sparse, but only in the sense that it leads to a few "large" topics, but the other topics may be non-negligible as well.)

To the best of our knowledge, the lemmas proven here were not derived elsewhere, so we include them for completeness.

For all of the claims below, we will be concerned with the following scenario:

$\vec{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ will be a vector of variables, and $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ a vector of parameters. We will let $\vec{\gamma}$ be distributed as $\vec{\gamma} := Dir(\alpha_1, \alpha_2, \ldots, \alpha_k)$, where $\alpha_i = C_i/k^c$, for some constants C_i and $c > 1$.

2.6.1 Sparsity

To characterize the sparsity of the topic proportions in a document, we will need the following lemma from (Telgarsky, 2013):

Lemma 51. *(Telgarsky, 2013) For a Dirichlet distribution with parameters $(C_1/k^c, C_2/k^c, \ldots, C_k/k^c)$, the probability that there are more than $c_0 \ln k$ coordinates in the Dirichlet draw that are $\geq 1/k^{c_0}$ is at most $1/k^{c_0}$.*

It's clear how this is related to our assumption: if one considers the coordinates $\geq \frac{1}{k^{c_0}}$ as "large", we assume, in a similar way, that there are only a few "large" coordinates. The difference is that we want the rest of the coordinates to be exactly zero.

2.6.2 Weak topic correlations

We will prove that the Dirichlet distribution satisfies something akin to the *weak topic correlations* property. We prove that when conditioning on some small ($o(k)$) set of topics being small, the marginal distributions for the rest of the topic proportions are very close to the original ones. This implies our "weak topic correlations" property.

The following is true:

Lemma 52. *Let $\vec{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ be distributed as specified above.*

Let S be a set of topics of size $o(k)$, and let's denote by γ_S the vector of variables corresponding to the topics in the set S, and $\gamma_{\bar{S}}$ the rest of the coordinates. Furthermore, let's denote by $\tilde{\gamma}_{\bar{S}}$ the distribution of $\gamma_{\bar{S}}$ conditioned on all the coordinates of γ_S being at most $1/k^{c_1}$ for $c_1 > 1$.

Then, for any $i \in \bar{S}$ and $\gamma = 1 - \delta$, any $\delta = \Omega(1)$,

$\mathbb{P}_{\gamma_S}(\gamma_i = \gamma) = (1 \pm o(1))\mathbb{P}_{\tilde{\gamma}_S}(\gamma_i = \gamma)$.

Proof. It's a folklore fact that if $\vec{Y} = \mathrm{Dir}(\vec{\alpha})$, then

$$(Y_1, Y_2, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_k | Y_i = y_i) = (1 - y_i)Dir(\alpha_1, \alpha_2, \ldots, \alpha_{i-1}, \alpha_{i+1}, \ldots, \alpha_k)$$

Applying this inductively, we get that $\tilde{\gamma}_{\bar{S}} = (1 - \sum_{j \in S} \gamma_i)\mathrm{Dir}(\vec{\alpha}_{\bar{S}})$. Let's denote $s := \sum_{j \in S} \gamma_i$, and $\tilde{s} = \sum_{i \in S} \alpha_i$. Then, since $\gamma_i \leq 1/k^{c_1}$ for $i \in S$, $s = o(1)$. Similarly, $\tilde{s} = o(1)$.

For notational convenience, let's call $\tilde{\alpha_0} = \sum_{i \notin S} \alpha_i$, and $\alpha_0 = \sum_i \alpha_i = \tilde{\alpha_0} + \tilde{s}$.

The marginal distribution of variable Y_i where $\vec{Y} = \mathrm{Dir}(\vec{\alpha})$ is $\mathrm{Beta}(\alpha_i, \alpha_0 - \alpha_i)$.

Hence,

$$\mathbb{P}_{\gamma_S}(\gamma_i = \gamma) = \frac{1}{B(\alpha_i, \tilde{\alpha_0} + \tilde{s} - \alpha_i)} \gamma^{\alpha_i - 1}(1 - \gamma)^{\tilde{\alpha_0} + \tilde{s} - \alpha_i - 1}$$

and

$$\mathbb{P}_{\tilde{\gamma}_{\bar{S}}}(\gamma_i = \gamma) = \frac{1}{B(\alpha_i, \tilde{\alpha_0} - \alpha_i)} (\frac{\gamma}{1 - s})^{\alpha_i - 1}(1 - \frac{\gamma}{1 - s})^{\tilde{\alpha_0} - \alpha_i - 1}$$

The following holds:

$$\frac{\gamma^{\alpha_i - 1}(1 - \gamma)^{\tilde{\alpha_0} + \tilde{s} - \alpha_i - 1}}{(\frac{\gamma}{1-s})^{\alpha_i - 1}(1 - \frac{\gamma}{1-s})^{\tilde{\alpha_0} - \alpha_i - 1}} =$$

$$(1 - s)^{\alpha_i - 1} \left(\frac{(1 - s)(1 - \gamma)}{1 - s - \gamma} \right)^{-\alpha_i - 1} (1 - \gamma)^{\tilde{s}} =$$

$$\left(1 + \frac{s}{1 - s - \gamma} \right)^{-\alpha_i - 1} (1 - \gamma)^{\tilde{s}}$$

Now, I claim the above expression is $1 \pm o(1)$.

We'll just prove this for each of the terms individually. Since $1 + \frac{s}{1-s-\gamma} \geq 1$ and $-1 - \alpha_i \leq -1$, it follows that $(1 + \frac{s}{1-s-\gamma})^{-\alpha_i - 1} \leq 1$. On the other hand, by Bernoulli's inequality, $(1 + \frac{s}{1-s-\gamma})^{-\alpha_i - 1} \geq 1 - (\alpha_i + 1)\frac{s}{1-s-\gamma} \geq 1 - o(1)$,

since $\gamma = 1 - \delta$, for some constant δ, by our assumptions.

For the second term, since $1 - \gamma \leq 1$ and $\tilde{s} \geq 0$, $(1-\gamma)^{\tilde{s}} \leq 1$. On the other hand, again by Bernoulli's inequality, $(1-\gamma)^{\tilde{s}} \geq 1 - \gamma\tilde{s} = 1 - o(1)$, as we needed.

Comparing $B(\alpha_i, \tilde{\alpha_0} + \tilde{s} - \alpha_i)$ and $B(\alpha_i, \tilde{\alpha_0} - \alpha_i)$ is not so much more difficult. By definition, $B(\alpha_i, \alpha_0 - \alpha_i) = \int_0^1 x^{\alpha_i - 1}(1-x)^{\alpha_0 - \alpha_i - 1}\, dx$, so

$$\frac{B(\alpha_i, \alpha_0 + \tilde{s} - \alpha_i)}{B(\alpha_i, \alpha_0 - \alpha_i)} =$$

$$\frac{\int_0^1 x^{\alpha_i - 1}(1-x)^{\tilde{\alpha_0} + \tilde{s} - \alpha_i - 1}\, dx}{\int_0^1 x^{\alpha_i - 1}(1-x)^{\tilde{\alpha_0} - \alpha_i - 1}\, dx}$$

We'll just bound each of the ratios

$$\frac{x^{\alpha_i - 1}(1-x)^{\tilde{\alpha_0} + \tilde{s} - \alpha_i - 1}}{x^{\alpha_i - 1}(1-x)^{\tilde{\alpha_0} - \alpha_i - 1}}$$

Namely, this is just $(1-x)^{\tilde{s}}$. Same as above, $1 - o(1) \leq (1-\gamma)^{\tilde{s}} \leq 1$. Hence, these are within a constant from each other.

□

2.6.3 Dominant topic equidistribution

Now, we pass to proving a smooth version of the dominant topic equidistribution property. Namely, for a threshold $x_0 = o(1)$, we can consider a topic "large" whenever it's bigger than x_0. We will show that for any topics Y_i, Y_j, the probabilities that $Y_i > x_0$ and $Y_j > x_0$ are within a constant from each other.

Mathematically formalizing the above statement, we will prove the following lemma:

Lemma 53. *Let $\vec{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ be distributed as specified above. Then, $\frac{\mathbb{P}(Y_i > x_0)}{\mathbb{P}(Y_j > x_0)} = O(1)$, for any i, j if $x_0 = o(1)$.*

Proof. As before, the marginal distribution of Y_i is Beta$(\alpha_i, \alpha_0 - \alpha_i)$. The Beta distribution pdf is just $\mathbb{P}(x) = \frac{x^{\alpha_i - 1}(1-x)^{\alpha_0 - \alpha_i - 1}}{B(\alpha_i, \alpha_0 - \alpha_i)}$, where $B(\alpha_i, \alpha_0 - \alpha_i) = \int_0^1 x^{\alpha_i - 1}(1-x)^{\alpha_0 - \alpha_i - 1}\, dx$.

Hence, the ratio we care about can be written as

$$\frac{(\int_{x_0}^1 x^{\alpha_i - 1}(1-x)^{\alpha_0 - \alpha_i - 1}\, dx)/B(\alpha_i, \alpha_0 - \alpha_i)}{(\int_{x_0}^1 x^{\alpha_j - 1}(1-x)^{\alpha_0 - \alpha_j - 1}\, dx)/B(\alpha_j, \alpha_0 - \alpha_j)}$$

To get a bound on this ratio, it's sufficient to bound the normalization constants $B(\alpha_i, \alpha_0 - \alpha_i)$ and $B(\alpha_j, \alpha_0 - \alpha_j)$, as well as the ratio $\frac{\int_{x_0}^1 x^{\alpha_i - 1}(1-x)^{\alpha_0 - \alpha_i - 1}\, dx}{\int_{x_0}^1 x^{\alpha_j - 1}(1-x)^{\alpha_0 - \alpha_j - 1}\, dx}$. Let's prove first that $B(\alpha_i, \alpha_0 - \alpha_i) \simeq \mathbb{B}(\alpha_j, \alpha_0 - \alpha_j)$

By definition, $B(\alpha_i, \alpha_0 - \alpha_i) = \int_0^1 x^{\alpha_i - 1}(1-x)^{\alpha_0 - \alpha_i - 1}\, dx$. The way we'll analyze this quantity is that we'll divide the integral in two parts, one from 0 to $\frac{1}{2}$ and one from $\frac{1}{2}$ to 1.

Since $\alpha_0 = O(1)$, it follows that $\alpha_0 - \alpha_i - 1 \gtrsim -1$ and $\alpha_0 - \alpha_i - 1 \lesssim 1$. Hence, $(1-x)^{\alpha_0 - \alpha_i - 1} = \Theta(1)$. It follows that

$$\int_0^{\frac{1}{2}} x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}\,dx \simeq \int_0^{\frac{1}{2}} x^{\alpha_i-1}\,dx =$$

$$\simeq \frac{(1/2)^{\alpha_i}}{\alpha_i} \simeq \frac{1}{\alpha_i}$$

where the last equality follows since $\frac{1}{2} \leq (1/2)^{\alpha_i} \leq 1$.

The second portion is not much more difficult. Since $\frac{1}{2} \leq \frac{1}{2}^{\alpha_i-1} \leq 1$, it follows

$$\int_{\frac{1}{2}}^1 x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}\,dx \simeq \int_{\frac{1}{2}}^1 (1-x)^{\alpha_0-\alpha_i-1}\,dx =$$

$$\simeq \frac{(1/2)^{\alpha_0-\alpha_i}}{\alpha_0-\alpha_i} \simeq \frac{1}{\alpha_0}$$

where the last two equalities come about since $-1 \lesssim \alpha_0 - \alpha_i \lesssim 1$.

But the above two estimates proved that for any i, $B(\alpha_i, \alpha_0 - \alpha_i) \simeq \frac{1}{\alpha_i}$, as we needed.

So, we proceed onto bounding

$$\frac{\int_{x_0}^1 x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}\,dx}{\int_{x_0}^1 x^{\alpha_j-1}(1-x)^{\alpha_0-\alpha_j-1}\,dx}$$

We'll proceed in a similar fashion as before. We'll pick some point x_T, and if $x < x_T$, we will show that $x^{\alpha_j-1}(1-x)^{\alpha_0-\alpha_j-1}$ is within a constant factor from $x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}$. On the other hand, we will show that part of the integral where $x > x_T$ is dominated by the part where $x < x_T$, which will imply the claim we need.

Let's rewrite the ratio above a little:

$$\frac{x^{\alpha_j-1}(1-x)^{\alpha_0-\alpha_j-1}}{x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}} =$$

$$(\frac{x}{1-x})^{\alpha_j-\alpha_i} = e^{(\alpha_j-\alpha_i)\ln(\frac{x}{1-x})}$$

Proceeding as outlined, I claim that for sufficiently large constants C_1, C_2, s.t. if $x \leq 1 - \frac{1}{1+C_1 e^{\frac{1}{\alpha_i}\frac{1}{C_2}}}$, then $\frac{x^{\alpha_j-1}(1-x)^{\alpha_0-\alpha_j-1}}{x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}} = O(1)$. Let's call $x_T = 1 - \frac{1}{1+C_1 e^{\frac{1}{\alpha_i}\frac{1}{C_2}}}$.

The claim is then, that if $x_T \geq x \geq x_0$, that $(\alpha_j - \alpha_i)\ln(\frac{x}{1-x}) = O(1)$.

First let's assume, $\alpha_j - \alpha_i \geq 0$.

Then, if $\ln(\frac{x}{1-x}) < 0 \Leftrightarrow x < \frac{1}{2}$, the condition is of course satisfied. So let's assume $x \geq \frac{1}{2}$. When $\frac{1}{2} \leq x \leq x_T$, we get that $\frac{x}{1-x} \leq C_1 e^{e^{\frac{1}{\alpha_j}\frac{1}{C_2}}}$. Hence, $\ln(\frac{x}{1-x}) \leq \ln C_1 + \frac{1}{\alpha_j}\frac{1}{C_2}$. It follows that if C_1, C_2 are sufficiently large,

$$(\frac{x}{1-x})^{\alpha_j-\alpha_i} \leq e^{\ln(\frac{x}{1-x})\alpha_j} = O(1)$$

On the other hand, if $\alpha_j - \alpha_i \leq 0$, when $x \geq \frac{1}{2}$, $(\alpha_j - \alpha_i)\ln(\frac{x}{1-x}) \leq 0$, so we are fine. However, since $|\alpha_j - \alpha_i| \leq \alpha_i$,

it's easy to check when $x \geq \frac{e^{-c_1/\alpha_i}}{1+e^{-c_1/\alpha_i}} > x_0$, that $(\alpha_j - \alpha_i)\ln(\frac{x}{1-x}) = O(1)$.

Finally, we want to claim that the portion of the integral from x_T to 1 is dominated by the portion from x_0 to x_T.

We can show that the latter portion is $O(e^{-k})$, and the first is $\Omega(1)$.

Let's lower bound the first portion. We lower bound $\int_{x_0}^{x_T} x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}\,dx$ by $x_T^{\alpha_i-1}\int_{x_0}^{x_T}(1-x)^{\alpha_0-\alpha_i-1}\,dx$. For the first factor in the above expression, we use Bernoulli's inequality to prove it's $\Omega(1)$. For the second, the integral will evaluate to

$$\frac{(1-x_0)^{\alpha_0-\alpha_i} - (1-x_T)^{\alpha_0-\alpha_i}}{\alpha_0-\alpha_i}$$

Let's lower bound the first term in the numerator. If $\alpha_0 - \alpha_i \geq 1$, another application of Bernoulli's inequality gives: $(1-x_0)^{\alpha_0-\alpha_i} \geq 1 - (\alpha_0-\alpha_i)x_0 \geq 1 - o(1)$. If, on the other hand, $0 \leq \alpha_0 - \alpha_i \leq 1$, $(1-x_0)^{\alpha_0-\alpha_i} \geq 1 - x_0 \geq 1 - o(1)$.

Then, I claim that $(1-x_T)^{\alpha_0-\alpha_i} = e^{-\Omega(k)}$. Indeed, for some constant C_3,

$$\left(\frac{1}{1+C_1 e^{\frac{1}{\alpha_j}\frac{1}{C_2}}}\right)^{\alpha_0-\alpha_i} \leq \left(\frac{1}{C_3 e^{\frac{1}{\alpha_j}\frac{1}{C_2}}}\right)^{\alpha_0-\alpha_i} =$$

$$= e^{-\ln(C_3 e^{\frac{1}{\alpha_j}\frac{1}{C_2}})(\alpha_0-\alpha_i)}$$

However, since $\alpha_0 = \Omega(K\alpha_j)$ and $\alpha_0 - \alpha_i = \Omega(\alpha_0)$, the above expression is upper bounded by $e^{-\Omega(k)}$, which is what we were claiming. Hence, $x_T^{\alpha_i-1}\int_{x_0}^{x_T}(1-x)^{\alpha_0-\alpha_i-1}\,dx = \Omega(1)$.

Let's upper bound the latter portion. This expression is upper bounded by

$$x_T^{\alpha_i-1}\int_{x_T}^{1}(1-x)^{\alpha_0-\alpha_i-1}\,dx = x_T^{\alpha_i-1}\frac{\left(\frac{1}{1+C_1 e^{\frac{1}{\alpha_j}\frac{1}{C_2}}}\right)^{\alpha_0-\alpha_i}}{\alpha_0-\alpha_i}$$

Now, we will separately bound each of $x_T^{\alpha_i-1}$ and $\frac{\left(\frac{1}{1+C_1 e^{\frac{1}{\alpha_j}\frac{1}{C_2}}}\right)^{\alpha_0-\alpha_i}}{\alpha_0-\alpha_i}$.

The first term can be written as $\frac{1}{x_T^{1-\alpha_i}}$. Now, since $1-\alpha_i \geq 0$, we can use Bernoulli's inequality to lower bound $x_T^{1-\alpha_i}$ by $1 - \frac{1}{1+C_1 e^{\frac{1}{\alpha_j}\frac{1}{C_2}}}(1-\alpha_i)$. Since $\frac{1}{1+C_1 e^{\frac{1}{\alpha_j}\frac{1}{C_2}}} = O(1/e^{\frac{1}{\alpha_j}})$, and $1-\alpha_i \leq 1/2$, let's say, $1 - \frac{1}{1+C_1 e^{\frac{1}{\alpha_j}\frac{1}{C_2}}}(1-\alpha_i) = \Omega(1)$, i.e. $x_T^{\alpha_i-1} = O(1)$.

For the second term, we already proved above that $(1-x_T)^{\alpha_0-\alpha_i} = e^{-\Omega(k)}$, This implies that $\int_{x_T}^{1} x^{\alpha_i-1}(1-x)^{\alpha_0-\alpha_i-1}\,dx = O(e^{-k})$, which finishes the proof.

□

2.6.4 Independent topic inclusion

Finally, there's a very simple proxy for "independent topic inclusion". Again, as above, $\tilde{\gamma}_{\bar{S}} = (1 - \sum_{j\in S}\gamma_i)\mathrm{Dir}(\vec{\alpha}_{\bar{S}})$.

But, if we consider "inclusion" the probability that a given topic is "noticeable" (i.e. $\geq \frac{1}{n^{c_0}}$, say), we can use the above Lemma 53 to show that the probability that any topic is "large" (but still $o(1)$) is within a constant for all the topics in $\bar{S}$.

2.7 Technical details: estimates on number of documents

Finally, we state a few helper lemmas to estimate how many documents will be needed. The properties we need are that the empirical marginals of a dominating topic in the documents where it's dominating are close to the actual ones, and similarly that the empirical marginals of the dominating topic, conditioned on the set of topics that a discriminative word belongs to not being present are close to the actual ones.

The former statement is the following:

Lemma 54. *Let $E_i = \mathbf{E}[\gamma^*_{d,i} | \gamma^*_{d,i}$ is dominating$]$. If the total number of documents is $D = \Omega(\frac{K \log^2 K}{\epsilon^2})$, and D_i is the number of documents where i is the dominant topic, then with high probability, for all topics i,*

$$(1-\epsilon)E_i \leq \frac{1}{D_i} \sum_{d \in D_i} \gamma^*_{d,i} \leq (1+\epsilon)E_i$$

Proof. Since documents are generated independently, $\Pr[\frac{1}{D_i} \sum_{d \in D_i} \gamma^*_{d,i} > (1+\epsilon)E_i] \leq e^{-\frac{\epsilon^2 D_i E_i}{3}}$ by Chernoff.

Since there are at most s topics per document, $E_i \geq \frac{1}{s}$, so $\Pr[\frac{1}{D_i} \sum_{d \in D_i} \gamma^*_{d,i} > (1+\epsilon)E_i] \leq e^{-\frac{\epsilon^2 D_i}{3T}}$

An analogous statement holds for $\Pr[\frac{1}{D_i} \sum_{d \in D_i} \gamma^*_{d,i} < (1-\epsilon)E_i]$

Then, if $D_i = \frac{\log^2 k}{\epsilon^2}$, by union bounding, we get that with high probability, for all topics, $(1-\epsilon)E_i \leq \frac{1}{D_i} \sum_{d \in D_i} \gamma^*_{d,i} \leq (1+\epsilon)E_i$

However, the probability of a topic being dominating is C_i/k for some constant C_i. So, by another Chernoff bound,

$$\Pr[D_i < (1-\epsilon)C_i D/k] \leq e^{-\frac{\epsilon^2 C_i D}{3k}} \tag{2.7.1}$$

So, if we take $D = \frac{k}{\epsilon^2} \log^2 k$, with high probability, for all topics, $D_i = \Theta(D/k)$.

Putting everything together, we get that if $D = \frac{k \log^2 k}{\epsilon^2}$, with high probability,

$$(1-\epsilon)E_i \leq \frac{1}{D_i} \sum_{d \in D_i} \gamma^*_{d,i} \leq (1+\epsilon)E_i$$

□

Next, we calculate how many documents are needed to match the marginals of the dominating topics, conditioned

on a small subset (of size $o(k)$) of the topics not being included in a document. More formally,

Lemma 55. *For the discriminative word j, let jS be the set of topics it belongs to. For a topic $i \in jS$, let Let $E_{i,jS} = \mathbf{E}[\gamma^*_{d,i} | \gamma^*_{d,i}$ is dominating, $\gamma^*_{d,i'} = 0, \forall i' \in jS]$. Let $D_{i,jS}$ be the number of documents where i is dominating, and $\gamma^*_{d,i'} = 0, \forall i' \in jS$.*

If the number of documents $D \geq \frac{k \log^2 n}{\epsilon^2}$, then with high probability, for all topics i and discriminative words j, $(1-\epsilon)E_{i,jS} \leq \frac{1}{D_{i,jS}} \sum_{d \in D_{i,jS}} \gamma^*_{d,i} \leq (1+\epsilon)E_{i,jS}$

Proof. Since $E_{i,jS} = (1 \pm o(1))E_i$, by the weak topic correlation property, an analogous proof as above shows that if we get that if $D_{i,jS} = \frac{\log^2 k}{\epsilon^2}$, with high probability, $(1-\epsilon)E_{iS} \leq \frac{1}{D_{iS}} \sum_{d \in D_{iS}} \gamma^*_{d,i} \leq (1+\epsilon)E_{iS}$.

But by the independent topic inclusion property, the probability of generating a document D with i being the dominating topic, s.t. no topics in jS appear in it is $\Theta(1/k)$. So, again by Chernoff,

$$\Pr[D_{i,jS} < (1-\epsilon)C_i D/k] \leq e^{-\frac{\epsilon^2 C_i D}{3k}} \tag{2.7.2}$$

If we take $D = \frac{k}{\epsilon^2} \log^2 n$, $\Pr[D_{i,jS} < (1-\epsilon)C_i D/k] \leq e^{-\log^2 n}$. However, since the total number of i, jS pairs is at most n^2, union bounding, we get that with high probability, for all pairs i, jS,

$$(1-\epsilon)E_{i,jS} \leq \frac{1}{D_{i,jS}} \sum_{d \in D_{i,jS}} \gamma^*_{d,i} \leq (1+\epsilon)E_{i,jS}$$

□

Finally, the following short lemma to estimate the number of documents in which a word j belongs only to the dominating topic is implicit in the proof above:

Lemma 56. *Let $D_{i,jS}$ be the number of documents where i is dominating, and $\gamma^*_{d,i'} = 0, \forall i' \in j_S$. If the number of documents $D \geq \frac{k \log^2 n}{\epsilon^2}$, then with high probability, for all topics i and discriminative words j, $D_{i,jS} \geq D_i(1-\epsilon)(1-o(1))$*

2.8 Changing the updates

In the prior sections, our goal was to analyze iterative updates which are as close as possible to the updates in variational Bayes updates used by practitioners. The reason for such a goal is to gain a better understanding of the structure of data sets where they can be expected to work. Of course, there is an obvious related goal to this: can we design an iterative algorithm of comparable runtime to variational Bayes, but with provable guarantees? In fact, we can even hope that such algorithms, inspired by theoretical considerations, can even do better in practice.

In this section, we do exactly that: we will take leave of the usual mean-field variational Bayes updates, and design a different algorithm, which also alternatively updates the topic-word matrix and the topic proportions estimates in the documents, and is of comparable computational efficiency. The benefit will be that we will be able to significantly relax the assumptions on the topic-word matrix and the requirements of the initialization, compared to the ones in the previous section. Additionally, the guarantee will be substantially more robust to noise in the data.

The slight cost of our change will be that the prior on the topic proportions we have to assume in the documents will have to be *completely independent* – so we will not be able to enforce the marginalization constraint on the topic proportions, i.e. $\sum_{i=1}^{k} \gamma_i = 1$. In that sense, the latent variable model we consider can be considered as a non-negative version of ICA. Nevertheless, we keep the notation of the previous section to delineate the close analogy to topic models. More concretely, we will consider the following latent-variable model:

$$\tilde{f} = \beta^* \gamma^* + \nu \tag{2.8.1}$$

where β^* is the topic-word matrix, γ^* are non-negative weights from a prior distribution, and ν is noise in the model. Similarly as before, our focus is to recover β^*, assuming some properties of β^*, x^*, and ν. The γ^* can be thought of as the topic proportions, thought we will not constrain them to sum to 1. More precisely, the assumptions are as follows. Let $[M]^i$ denote the i-th row of a matrix M, $[M]_j$ its i-th column, $M_{i,j}$ its (i,j)-th entry. Denote its column norm, row norm, and symmetrized norm as $\|M\|_1 = \max_j \sum_i \left|M_{i,j}\right|, \|M\|_\infty = \max_i \sum_j \left|M_{i,j}\right|$, and $\|M\|_{\mathrm{s}} = \max\left\{\|M\|_1, \|M\|_\infty\right\}$, respectively. We assume the following hold for parameters $C_1, c_2, C_2, \ell, C_\nu$ to be determined in our theorems.

(**A1**) The columns of β^* are linearly independent.

(**A2**) For all $i \in [k]$, $\gamma_i^* \in [0,1]$, $\mathbb{E}[\gamma_i^*] \leq \frac{C_1}{k}$ and $\frac{c_2}{k} \leq \mathbb{E}[(\gamma_i^*)^2] \leq \frac{C_2}{k}$, and γ_i^*'s are independent.

(**A3**) The initialization $\beta^0 = \beta^*(\Sigma^0 + E^0) + N^{(0)}$, where Σ^0 is diagonal, E^0 is off-diagonal, and

$$\Sigma^0 \geq (1-\ell)\mathrm{Id}, \quad \left\|E^0\right\|_{\mathrm{s}} \leq \ell.$$

Furthermore, we consider two noise models.

(**N1**) Adversarial noise: only assume that $\max_i |\nu_i| \leq C_\nu$ almost surely.

(**N2**) Unbiased noise: $\max_i |\nu_i| \leq C_\nu$ almost surely, and $\mathbb{E}[\nu | x^*] = 0$.

Several remarks about the assumptions are in order. (**A1**) significantly relaxes the assumptions of the previous section – and is needed to ensure identifiability. Otherwise, for instance, if $(\beta^*)_3 = \lambda_1(\beta^*)_1 + \lambda_2(\beta^*)_2$, it is impossible

to distinguish between the case when $\gamma_3^* = 1$ and the case when $\gamma_2^* = \lambda_1$ and $\gamma_1^* = \lambda_2$. Note that, in principle, we do not restrict the feature matrix to be non-negative even! This is helpful if instead of topic modeling, we think of this generative model as one in which the samples $\tilde{f}$ are non-negative combinations of "features" specified by the columns of β^*.

(**A2**) is essentially where our model differs from topic modeling: the coordinates of γ need to be independent. In this sense, this model is like a non-negative variant of ICA. Apart from that, (**A2**) constraints the coordinates to be non-negative and bounded by 1; this is simply a matter of scaling. The second moment assumptions ensure that, roughly speaking, each topic appears with reasonable probability. This is expected: if the occurrences of the topics are extremely unbalanced, then it will be difficult to recover the rare ones.

The warm start required is also substantially relaxed from the previous section. (**A3**) specifies that each feature β_i^0 has a large fraction of the ground-truth feature β_i^* and a small fraction of the other features, plus some noise outside the span of the ground-truth features. Importantly, unlike the previous section, this fraction *does not* depend on the range of the entries in β^*, nor does it require anchor words. Moreover, we can tolerate $N^{(0)}$: a component of β^0 outside the column space of β^*.

The adversarial noise model (**N1**) is very general, only imposing an upper bound on the entry-wise noise level. Thus, ν can be correlated with γ^* in some complicated unknown way. (**N2**) additionally requires it to be zero mean, which is commonly assumed and will be exploited by our algorithm to tolerate larger noise.

2.8.1 The new updates: main algorithm

In this section we outline the main algorithm, and explain the basic intuition behind the proof of correctness of these modified updates.

Algorithm 6 Purification

Input: initialization β^0, threshold α, step size η, scaling factor r, sample size D, iterations T

1: **for** $t = 0, 1, 2, ..., T-1$ **do**

2: Draw examples $y_1, \ldots, y_D$.

3: (Decode) Compute $\beta^\dagger$, the pseudo-inverse of β^t with minimum $\|(\beta)^\dagger\|_\infty$.

Set $x = \phi_\alpha(\beta^\dagger y)$ for each example y. *// ϕ_α is ReLU activation; see (2.8.2) for the definition*

4: (Update) Update the feature matrix

$$\beta^{t+1} = (1-\eta)\beta^t + r\eta\hat{\mathbb{E}}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right]$$

where $\hat{\mathbb{E}}$ is over independent uniform $\tilde{f}, \tilde{f}'$ from $\left\{\tilde{f}_1, \ldots, \tilde{f}_D\right\}$, and γ, γ' are their decodings.

5: **end for**

Output: $\beta = \beta^{(T)}$

The main algorithm is delineated as Algorithm 6. It keeps a working topic-word matrix and operates in iterations. In each iteration, it first compute the weights for a batch of D examples (*decoding*), and then uses the computed weights to update the feature matrix (*updating*).

The decoding is simply multiplying the example by the pseudo-inverse of the current feature matrix and then

passing it through the rectified linear unit (ReLU) ϕ_α with offset α. The pseudo-inverse with minimum infinity norm is used so as to maximize the robustness to noise (see the theorems). The ReLU function ϕ_α operates element-wise on the input vector v, and for an element v_i, it is defined as

$$\phi_\alpha(v_i) = \max\{v_i - \alpha, 0\}. \tag{2.8.2}$$

To get an intuition why the decoding makes sense, suppose the current feature matrix is the ground-truth. Then $\beta^\dagger \tilde{f} = \beta^\dagger \beta^* \gamma^* + \beta^\dagger \nu = \gamma^* + \beta^\dagger \nu$. So we would like to use a small $\beta^\dagger$ and use threshold to remove the noise term.

In the encoding step, the algorithm move the feature matrix along the direction $\mathbb{E}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right]$. To see intuitively why this is a good direction, note that when the decoding is perfect and there is no noise, $\mathbb{E}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right] = \beta^*$, and thus it is moving towards the ground-truth. Without those ideal conditions, we need to choose a proper step size, which is tuned by the parameters η and r.

2.8.2 Results for a simplified case

In order to demonstrate the intuition and results clearly, we first consider a simplified setting first, with assumptions (**A1**), (**A2'**), (**A3**), and (**N1**), where

(**A2'**) γ_i^*'s are independent, and $\gamma_i^* = 1$ with probability s/n and 0 otherwise for a constant $s > 0$.

Furthermore, we will assume $N^0 = 0$.

Note this is a special case of our general assumptions, with $C_1 = c_2 = C_2 = s$ where s is the parameter in (**A2'**). Moreover, this is quite close to the setting we considered in Section 2.2. We will subsequently present the general result in Section 2.8.5, which will be hopefully easier to digest after we have presented this simpler setting.

For notational convenience, let $(\beta^*)^\dagger$ denote the matrix satisfying $(\beta^*)^\dagger \beta^* = \mathrm{Id}$. If there are multiple such matrices we let it denote the one with minimum $\|(\beta^*)^\dagger\|_\infty$.

Theorem 57 (Simplified case, adversarial noise, (Li et al., 2016))**.** *There exists an* absolute constant $\mathcal{G}$ *such that if Assumptions (**A1**),(**A2'**),(**A3**) and (**N1**) are satisfied with $l = 1/10$, $C_\nu \le \frac{\mathcal{G}c}{\max\{n,k\|(\beta^*)^\dagger\|_\infty\}}$ for some $0 \le c \le 1$ and $N^0 = 0$, then there is a choice of parameters α, η, r such that for every $0 < \epsilon, \delta < 1$ and $D = \mathrm{poly}(k, n, 1/\epsilon, 1/\delta)$ the following holds with probability at least $1 - \delta$:*

After $T = O\left(\ln \frac{1}{\epsilon}\right)$ iterations, Algorithm 6 outputs a solution $\beta = \beta^*(\Sigma + E) + N$ where $\Sigma \ge (1 - \ell)\mathrm{Id}$ is diagonal, $\|E\|_1 \le \epsilon + c$ is off-diagonal, and $\|N\|_1 \le c$.

Remarks. Consequently, when $\|\beta^*\|_1 = 1$, we can do normalization $\hat{\beta}_i = \beta_i / \|\beta_i\|_1$, and the normalized output $\hat{\beta}$ satisfies

$$\|\hat{\beta} - \beta^*\|_1 \le \epsilon + 2c.$$

In particular, under mild conditions and with proper parameters, our algorithm recovers the ground-truth in a geometric rate. It can achieve arbitrary small recovery error in the noiseless setting, and achieve error up to the noise limit even with adversarial noise whose level is comparable to the signal.

The result implies that with large adversarial noise, the algorithm can still recover the features up to the noise limit. When $n \geq k\|(\beta^*)^\dagger\|_\infty$, each data point has adversarial noise with ℓ_1 norm as large as $\|\nu\|_1 = C_\nu n = \Omega(c)$, which is in the same order as the signal $\|\beta^*\gamma^*\|_1 = O(1)$. Our algorithm still works in this regime. Furthermore, the *final* error $\|\beta - \beta^*\|_1$ is $O(c)$, in the same order as the *adversarial* noise in *one* data point.

Note the appearance of $\|(\beta^*)^\dagger\|_\infty$ is not surprising. The case when the columns are the canonical unit vectors for instance, which corresponds to $\|(\beta^*)^\dagger\|_\infty = 1$, is expected to be easier than the case when the columns are nearly the same, which corresponds to large $\|(\beta^*)^\dagger\|_\infty$.

A similar theorem holds for the unbiased noise model.

Theorem 58 (Simplified case, unbiased noise, (Li et al., 2016))**.** *If Assumptions **(A1)**,**(A2')**,**(A3)** and **(N2)** are satisfied with* $C_\nu = \frac{\mathcal{G}c\sqrt{k}}{\max\{n,k\|(\beta^*)^\dagger\|_\infty\}}$*, then the same guarantee as Theorem 57 holds.*

Remarks. With unbiased noise which is commonly assumed in many applications, the algorithm can tolerate noise level $\sqrt{k}$ larger than the adversarial case. When $n \geq k\|(\beta^*)^\dagger\|_\infty$, each data point has noise with ℓ_1 norm as large as $\|\nu\|_1 = C_\nu n = \Omega(c\sqrt{k})$, which can be $\Omega(\sqrt{k})$ times larger than the signal $\|\beta^*\gamma^*\|_1 = O(1)$. The algorithm can recover the ground-truth in this heavy noise regime. Furthermore, the *final* error $\|\beta - \beta^*\|_1$ is $O\left(\|\nu\|_1/\sqrt{k}\right)$, which is only $O(1/\sqrt{k})$ fraction of the noise in *one* data point. This is a strong denoising effect and a bit counter-intuitive. It is possible since we exploit averaging of the noise for cancellation, as well as thresholding to remove noise spread out in the coordinates.

2.8.3 Analysis: intuition

A natural approach typically employed to analyze algorithms for non-convex problems is to define a function on the intermediate solution β and the ground-truth β^* measuring their distance and then show that the function decreases at each step. However, a single potential function will not be enough in our case, as we argue below, so we introduce a novel framework of maintaining two potential functions which capture different aspects of the intermediate solutions.

Let us denote the intermediate solution and the update as (omitting the superscript (t))

$$\beta = \beta^*(\Sigma + E) + N, \quad \hat{\mathbb{E}}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top] = \beta^*(\widetilde{\Sigma} + \widetilde{E}) + \widetilde{N}, \tag{2.8.3}$$

where Σ and $\widetilde{\Sigma}$ are diagonal, E and $\widetilde{E}$ are off-diagonal, and N and $\widetilde{N}$ are the terms outside the span of β^* which is caused by the noise. To cleanly illustrate the intuition behind ReLU and the coupled potential functions, we focus on

the noiseless case and assume that we have infinite samples.

Since $\beta_i = \Sigma_{i,i}\beta_i^* + \sum_{j\neq i} E_{j,i}\beta_j^*$, if the ratio between $\|E_i\|_1 = \sum_{j\neq i}\left|E_{j,i}\right|$ and $\Sigma_{i,i}$ gets smaller, then the algorithm is making progress; if the ratio is large at the end, a normalization of β_i gives a good approximation of β_i^*. So it suffices to show that $\Sigma_{i,i}$ is always about a constant while $\|E_i\|_1$ decreases at each iteration. We will focus on E and consider the update rule in more detail to argue this. After some calculation, we have

$$E \leftarrow (1-\eta)E + r\eta\widetilde{E}, \qquad \widetilde{E} = \mathbb{E}[(\gamma^* - (\gamma')^*)(\gamma - \gamma')^\top], \tag{2.8.4}$$

where γ, γ' are the decoding for $\gamma^*, (\gamma')^*$ respectively:

$$\gamma = \phi_\alpha\left((\Sigma+E)^{-1}\gamma^*\right), \qquad \gamma' = \phi_\alpha\left((\Sigma+E)^{-1}(\gamma')^*\right). \tag{2.8.5}$$

To see why the ReLU function matters, consider the case when we do not use it.

$$\begin{aligned}\widetilde{E} &= \mathbb{E}(\gamma^* - (\gamma')^*)\left[\beta^\dagger\beta^*(\gamma^* - (\gamma')^*)\right]^\top = \mathbb{E}\left[(\gamma^* - (\gamma')^*)(\gamma^* - (\gamma')^*)^\top\right]\left[(\Sigma+E)^{-1}\right]^\top \\ &\propto \left[(\Sigma+E)^{-1}\right]^\top \approx \Sigma^{-1} - \Sigma^{-1}E\Sigma^{-1}.\end{aligned}$$

where we used Taylor expansion and the fact that $\mathbb{E}\left[(\gamma^* - (\gamma')^*)(\gamma^* - (\gamma')^*)^\top\right]$ is a scaling of identity. Hence, if we think of Σ as approximately Id and take an appropriate r, the update to the matrix E is approximately $E \leftarrow E - \eta E^\top$. Since we do not have control over the signs of E throughout the iterations, the problematic case is when the entries of $E^\top$ and E roughly match in signs, which would lead to the entries of E increasing.

Now we consider the decoding to see why the ReLU is helpful. Ignoring the higher order terms and regarding $\Sigma = \mathrm{Id}$, we have

$$\gamma = \phi_\alpha\left((\Sigma+E)^{-1}\gamma^*\right) \approx \phi_\alpha\left(\Sigma^{-1}\gamma^* - \Sigma^{-1}E\Sigma^{-1}\gamma^*\right) \approx \phi_\alpha\left(\gamma^* - E\gamma^*\right). \tag{2.8.6}$$

The problematic term is $E\gamma^*$. These errors when summed up will be comparable or even larger than the signal, and the algorithm will fail. However, since the signal coordinates are non-negative and most coordinates with errors only have small values, the hope is that thresholding with ReLU can remove those errors while keeping a large fraction of the signal coordinates. This leads to large $\widetilde{\Sigma}_{i,i}$ and small $\widetilde{E}_{j,i}$'s, and then we can choose an r such that $E_{j,i}$'s keep decreasing while $\Sigma_{i,i}$'s stay in a certain range.

To quantify the intuition above, we need to divide E into its positive part E_+ and its negative part E_-:

$$[E_+]_{i,j} = \max\left\{E_{i,j}, 0\right\}, \qquad [E_-]_{i,j} = \max\left\{-E_{i,j}, 0\right\}. \tag{2.8.7}$$

The reason to do so is the following: when $E_{i,j}$ is negative, by the Taylor expansion approximation, $\left[(\Sigma+E)^{-1}\gamma^*\right]_i$ will tend to be more positive and will not be thresholded to 0 by the ReLU most of the time. Therefore, $E_{j,i}$ will become more positive at next iteration. On the other hand, when $E_{i,j}$ is positive, $\left[(\Sigma+E)^{-1}\gamma^*\right]_i$ will tend to be more negative and zeroed out by the ReLU function. Therefore, $E_{j,i}$ will *not* be more negative at next iteration. Informally, we will show for positive and negative parts of E:

$$\text{postive}^{t+1} \leftarrow (1-\eta)\text{positive}^t + (\eta)\text{negative}^t, \text{negative}^{t+1} \leftarrow (1-\eta)\text{negative}^t + (\varepsilon\eta)\text{positive}^t$$

for a small $\varepsilon \ll 1$. Due to the appearance of ε in the above updates, we can "couple" the two parts, namely show that a weighted average of them will decrease, which implies that $\|E\|_s$ is small at the end. This leads to our coupled potential function.[4]

2.8.4 Analysis: proof sketch

We now provide a proof sketch for the simplified case presented above. The complete proof of the results for the general case (which is stated in the next section) is presented in the appendix. The lemmas here are direct corollaries of those in the appendix.

One iteration. We focus on one update and omit the superscript t. Recall the definitions of E, Σ, N and $\widetilde{E}$, $\widetilde{\Sigma}$ and $\widetilde{N}$ from (2.8.3). Our goal is to derive lower and upper bounds for $\widetilde{E}$, $\widetilde{\Sigma}$ and $\widetilde{N}$, assuming that $\Sigma_{i,i}$ falls into some range around 1, while E and N are small. This will allow us to do induction on t.

First, begin with the decoding. A simple calculation shows that the decoding for $\tilde{f} = \beta^*\gamma^* + \nu$ is

$$\gamma = \phi_\alpha\left(Z\gamma^* + \xi\right), \quad \text{where } Z = (\Sigma+E)^{-1}, \ \xi = -\beta^\dagger NZ\gamma^* + \beta^\dagger\nu. \tag{2.8.8}$$

Now, we can present our key lemmas bounding $\widetilde{E}$, $\widetilde{\Sigma}$, and $\widetilde{N}$. Before doing this, we add that the particular value for r we will choose is $r = \frac{n}{s}$ (recalling s is the sparsity of γ^* according to Assumption (**A2'**)). We also set the threshold of the ReLU as $\rho < \alpha \ll \frac{s}{n}$. Then, we get:

Lemma 59 (Simplified bound on $\widetilde{E}$, informal)**.** *(1) if $Z_{i,j} < 0$, then $\left|\widetilde{E}_{j,i}\right| \le o\left(\frac{s}{n}\left(|Z_{i,j}| + \rho\right)\right)$,*
(2) if $Z_{i,j} \ge 0$, then $-O\left(\left(\frac{s}{n}\right)^2 Z_{i,j} + \rho Z_{i,j}\right) \le \widetilde{E}_{j,i} \le O\left((\frac{s}{n}+\rho)|Z_{i,j}|\right)$.

Note that $Z \approx \Sigma^{-1} - \Sigma^{-1}E\Sigma^{-1}$, so $Z_{i,j} < 0$ corresponds roughly to $E_{i,j} > 0$. In this case, keeping in mind that $r = \frac{n}{s}$, the upper bound on $|\widetilde{E}_{j,i}|$ is small enough to ensure $|E_{j,i}|$ decreases, as described in the intuition.

On the other hand, when $Z_{i,j} \ge 0$ (roughly $E_{i,j} < 0$), the upper bound on $\widetilde{E}_{j,i}$ is large enough that $r\widetilde{E}_{j,i}$ can be

[4] Note that since intuitively, $E_{i,j}$ gets affected by $E_{j,i}$ after an update, if we have a row which contains negative entries, it is possible that $\|\beta_i - \beta_i^*\|_1$ increases. So we cannot simply use $\max_i \|\beta_i - \beta_i^*\|_1$ as a potential function.

on the same order as $E_{i,j}$, corresponding to the intuition that negative $E_{i,j}$ can contribute a large positive value to $E_{j,i}$. Fortunately, the lower bound on $\widetilde{E}_{j,i}$ is of much smaller absolute value, which allows us to show that a potential function that couples Case (1) and Case (2) in Lemma 59 actually decreases; see the induction below.

Lemma 60 (Simplified bound on $\widetilde{\Sigma}$, informal)**.** $\widetilde{\Sigma}_{i,i} \geq \Omega((\Sigma_{i,i}^{-1} - \alpha)/n)$.

Lemma 61 (Simplified bound on $\widetilde{N}$, adversarial noise, informal)**.** $\left|\widetilde{N}_{i,j}\right| \leq O(C_\nu/n)$.

Induction by iterations. We now show how to use the three lemmas to prove the theorem for the adversarial noise. The proof for the unbiased noise statement is similar.

Let $a_t := \left\|E_+^t\right\|_{\mathrm{s}}$ and $b_t := \left\|E_-^t\right\|_{\mathrm{s}}$, and choose $\eta = \ell/6$. We begin with proving the following three claims by induction on t: at the beginning of iteration t,

(1) $(1-\ell)\mathrm{Id} \preceq \Sigma^t$

(2) $\left\|E^t\right\|_{\mathrm{s}} \leq 1/8$, and if $t > 0$, then $a_t + \beta b_t \leq \left(1 - \frac{1}{25}\eta\right)(a_{t-1} + \beta b_{t-1}) + \eta h$, for some $\beta \in (1, 8)$, and some small value h,

(3) $\left\|N^t\right\|_{\mathrm{s}} \leq c/10$.

The most interesting part is the second claim. At a high level, by Lemma 59, we can show that

$$a_{t+1} \leq \left(1 - \frac{3}{25}\eta\right)a_t + 7\eta b_t + \eta h, \qquad b_{t+1} \leq \left(1 - \frac{24}{25}\eta\right)b_t + \frac{1}{100}\eta a_t + \eta h.$$

Notice that the contribution of b_t to a_{t+1} is quite large (due to the larger upper bound in Case (2) in Lemma 59), but the other contributions are all small. This allows to choose a $\beta \in (1, 8)$ so that $a_{t+1} + \beta b_{t+1}$ leads to the desired recurrence in the second claim. In other words, $a_{t+1} + \beta b_{t+1}$ is our potential function which decreases at each iteration up to the level h. The other claims can also be proved by the corresponding lemmas. Then the theorem follows from the induction claims.

2.8.5 More general results

More general weight distributions. Our argument holds under more general assumptions on x^*.

Theorem 62 (Adversarial noise, (Li et al., 2016))**.** *There exists an absolute constant $\mathcal{G}$ such that if Assumption **(A0)**-**(A3)** and **(N1)** are satisfied with $l = 1/10$, $C_2 \leq 2c_2$, $C_1^3 \leq \mathcal{G}c_2^2 n$, $C_\nu \leq \left\{\frac{c_2^2\mathcal{G}c}{C_1^2 m}, \frac{c_2^4\mathcal{G}c}{C_1^5 k\left\|(\beta^*)^\dagger\right\|_\infty}\right\}$ for $0 \leq c \leq 1$, and $\left\|N^0\right\|_\infty \leq \frac{c_2^2\mathcal{G}c}{C_1^3\left\|(\beta^*)^\dagger\right\|_\infty}$, then there is a choice of parameters α, η, r such that for every $0 < \epsilon, \delta < 1$ and $N = \mathrm{poly}(k, n, 1/\epsilon, 1/\delta)$, with probability at least $1 - \delta$ the following holds:*

After $T = O\left(\ln \frac{1}{\epsilon}\right)$ iterations, Algorithm 6 outputs a solution $\beta = \beta^*(\Sigma + E) + N$ where $\Sigma \succeq (1-\ell)\mathrm{Id}$ is diagonal, $\|E\|_1 \leq \epsilon + c/2$ is off-diagonal, and $\|N\|_1 \leq c/2$.

Theorem 63 (Unbiased noise, (Li et al., 2016)). *If Assumption (A0)-(A3) and (N2) are satisfied with* $C_\nu = \frac{c_2\mathcal{G}\sqrt{ck}}{C_1 \max\{n,k\}\|(\beta^*)^\dagger\|_\infty\}}$ *and the other parameters set as in Theorem 62, then the same guarantee holds.*

The conditions on C_1, c_2, C_2 intuitively mean that each feature needs to appear with reasonable probability. $C_2 \le 2c_2$ means that their proportions are reasonably balanced. This may be a mild restriction for some applications – however, we additionally propose a pre-processing step that can relax this in the following subsection.

The conditions allow a rather general family of distributions, so we point out an important special case to provide a more concrete sense of the parameters. For example, for the uniform independent distribution considered in the simplified case, we can actually allow s to be much larger than a constant; our algorithm just requires $s \le \mathcal{G}k$ for a fixed constant $\mathcal{G}$. So it works for uniform sparse distributions even when the sparsity is linear, which is an order of magnitude larger than what can be achieved in the dictionary learning regime. Furthermore, the distributions of γ_i^* can be very different, since we only require $C_1^3 = O(c_2^2 k)$. Moreover, all these can be handled without specific structural assumptions on β^*.

More general proportions. A mild restriction in Theorem 62 and 63 is that $C_2 \le 2c_2$, that is, $\max_{i\in[k]} \mathbb{E}[(\gamma_i^*)^2] \le 2\min_{i\in[k]} \mathbb{E}[(\gamma_i^*)^2]$. To relax this, we propose a pre-processing algorithm for balancing $\mathbb{E}[(\gamma_i^*)^2]$.

The idea is quite natural: instead of solving $\tilde{f} \approx \beta^*\gamma^*$, we could also solve $\tilde{f} \approx [\beta^* D][(D)^{-1}\gamma^*]$ for a positive diagonal matrix D, where $\mathbb{E}[(\gamma_i^*)^2]/D_{i,i}^2$ is with in a factor of 2 from each other. We show in the appendix that this can be done under assumptions as the above theorems, and additionally $\Sigma \preceq (1+\ell)\mathrm{Id}$ and $E^0 \ge 0$ entry-wise. After balancing, one can use Algorithm 6 on the new ground-truth matrix $[\beta^* D]$ to get the final result.

2.8.6 Technical details: proof of correctness of main algorithm

In this section, we formally prove the statement of Theorem 62 and 63. The way we proceed is to first analyze one update step, bounding the changes of Σ, E, N and some auxiliary variables, and then in the next subsection we put things together to prove the theorem.

2.8.6.1 Analysis of one update step

In the interest of readability, throughout this subsection we will focus on a particular iteration t and omit the superscript (t), while in the next subsection we will put back the superscript. For analysis, denote $\beta^{(t)}$ as

$$\beta = \beta^*(\Sigma + E) + N$$

where Σ is a diagonal matrix, E is an off-diagonal matrix, and N is the component of β that lies outside the span of β^* (e.g., the noise caused by the noise in the sample).

Recall the following notation:

$$
\begin{aligned}
Z &= (\Sigma + E)^{-1}, \\
V &= Z - \Sigma^{-1} = \Sigma^{-1} \sum_{k=1}^{\infty} (-E\Sigma^{-1})^k, \\
\xi &= -\beta^\dagger N Z x^* + \beta^\dagger \nu.
\end{aligned}
$$

Consider the update term $\hat{\mathbb{E}}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right]$ and denote it as

$$
\Delta = \hat{\mathbb{E}}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right] = \beta^*(\widetilde{\Sigma} + \widetilde{E}) + \widetilde{N}
$$

where $\widetilde{\Sigma}$ is a diagonal matrix, $\widetilde{E}$ is an off-diagonal matrix, and N is the component of Δ that lies outside the span of β^*.

Since we now use empirical average, we will have sampling noise. Denote it as

$$
N_s = \hat{\mathbb{E}}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top] - \mathbb{E}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top].
$$

Then by definition, for $\tilde{f} = \beta^*\gamma^* + \nu$ and $\tilde{f}' = \beta^*(\gamma')^* + \nu'$, we have

$$
\begin{aligned}
\hat{\mathbb{E}}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top] &= \mathbb{E}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top] + N_s \\
&= \beta^* \underbrace{\mathbb{E}\left[(\gamma^* - (\gamma')^*)(\gamma - \gamma')^\top\right]}_{\widetilde{\Sigma}+\widetilde{E}} + \underbrace{\mathbb{E}\left[(\nu - \nu')(\gamma - \gamma')^\top\right] + N_s}_{\widetilde{N}}.
\end{aligned}
$$

Our goal is then bounding $\widetilde{\Sigma}, \widetilde{E}, \widetilde{N}$ in terms of Σ, E, N. Before doing so, we present a lemma for the decoding.

Lemma 64 (Main: Decoding)**.** *Let $n \geq k$ be two positive integers. Let $\beta \in \mathbb{R}^{n\times k}$ be a matrix such that $\beta = \beta^*(\Sigma + E) + N$ where β^* is full rank, Σ is a diagonal matrix such that $\Sigma \geq \frac{1}{2}\mathrm{Id}$ and $\|E\|_1 < \frac{1}{2}$. Then for $\tilde{f} = \beta^*\gamma^* + \nu$, the decoding is*

$$
\begin{aligned}
\gamma &= \phi_\alpha\left(Z\gamma^* + \xi\right) \\
&= \phi_\alpha\left(\left(\Sigma^{-1} + V\right)\gamma^* + \xi\right).
\end{aligned}
$$

Proof of Lemma 64. Since $\beta = \beta^*(\Sigma + E) + N$, we have

$$
\begin{aligned}
\beta^* &= (\beta - N)(\Sigma + E)^{-1} \\
\tilde{f} &= (\beta - N)(\Sigma + E)^{-1}\gamma^* + \nu.
\end{aligned}
$$

Plugging into the decoding we get the first statement.

Observing that $\Sigma + E = (\mathrm{Id} + E\Sigma^{-1})\Sigma$ and $\|E\Sigma^{-1}\|_1 \le \|\Sigma^{-1}\|_1\|E\|_1 \le 2\|E\|_1 < 1$, we have $(\Sigma + E)^{-1} = \left(\Sigma^{-1} + V\right)$, resulting in the second statement. □

Lemma 65 (Main: Bound on $\widetilde{\Sigma}$)**.** *Suppose $|\xi_i| \le \rho < \alpha$ for any example and every $i \in [k]$, and suppose $\Sigma \ge \frac{1}{2}\mathrm{Id}$. Then for any $i \in [k]$,*

$$\widetilde{\Sigma}_{i,i} \ge \mathbb{E}\left[(\gamma_i^*)^2\right]\left(2\Sigma_{i,i}^{-1} - 2\left|V_{i,i}\right|\right) - \frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}\Sigma_{i,i}^{-1} + \frac{2C_1}{k}\left\|[V]^i\right\|_1\right),$$
$$\widetilde{\Sigma}_{i,i} \le \mathbb{E}\left[(\gamma_i^*)^2\right]\left(2\Sigma_{i,i}^{-1} + 2\left|V_{i,i}\right|\right) + \frac{2C_1}{k}\left(\rho + \frac{C_1}{k}\left\|[V]^i\right\|_1\right).$$

Proof of Lemma 65. According to the definition, we have

$$\begin{aligned}\widetilde{\Sigma}_{i,i} &= \left[(\beta^*)^\dagger \mathbb{E}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top]\right]_{i,i} \\ &= \mathbb{E}\left[(\gamma_i^* - (\gamma_i')^*)(\gamma_i - \gamma_i')\right] \\ &= \mathbb{E}\left[(\gamma_i^* - (\gamma_i')^*)\gamma_i\right] + \mathbb{E}\left[((\gamma_i')^* - \gamma_i^*)\gamma_i'\right].\end{aligned}$$

Since $(\gamma_i^* - (\gamma_i')^*)\gamma_i$ and $((\gamma_i')^* - \gamma_i^*)\gamma_i'$ has the same distribution, and $(\gamma')^*, \gamma^*$ are i.i.d. , we have

$$\begin{aligned}\widetilde{\Sigma}_{i,i} &= 2\mathbb{E}\left[(\gamma_i^* - (\gamma_i')^*)\gamma_i\right] \\ &= 2\mathbb{E}[\gamma_i^*\gamma_i] - 2\mathbb{E}[\gamma_i^*]\mathbb{E}[\gamma_i].\end{aligned}$$

So it suffices to bound $\mathbb{E}[\gamma_i^*\gamma_i]$ and $\mathbb{E}[\gamma_i]$. To do so, we first take a look at γ_i. By the decoding rule,

$$\gamma_i = \left[\phi_\alpha\left(\left(\Sigma^{-1} + V\right)\gamma^* + \xi\right)\right]_i.$$

Since ϕ_α is 1-Lipschitz, denoting $\Delta = \left|[V\gamma^*]_i + \xi_i\right|$ we have

$$\left[\phi_\alpha\left(\Sigma^{-1}\gamma^*\right)\right]_i - \Delta \le \gamma_i \le \left[\phi_\alpha\left(\Sigma^{-1}\gamma^*\right)\right]_i + \Delta. \tag{2.8.9}$$

For $\left[\phi_\alpha\left(\Sigma^{-1}\gamma^*\right)\right]_i$, by the Property **??** of $\phi_\alpha(z)$,

$$\Sigma_{i,i}^{-1}\gamma_i^* - \alpha \le \left[\phi_\alpha\left(\Sigma^{-1}\gamma^*\right)\right]_i = \phi_\alpha\left(\Sigma_{i,i}^{-1}\gamma_i^*\right) \le \Sigma_{i,i}^{-1}\gamma_i^*. \tag{2.8.10}$$

For $\Delta = \left|[V\gamma^*]_i + \xi_i\right|$,

$$\begin{aligned}
\mathbb{E}[\Delta] &\le \mathbb{E}\left[\left|\sum_j V_{i,j}\gamma_j^*\right|\right] + \mathbb{E}\left[|\xi_i|\right] \\
&\le \mathbb{E}\left[\sum_j \left|V_{i,j}\right|\gamma_j^*\right] + \rho \\
&= \sum_j \left|V_{i,j}\right|\mathbb{E}\left[\gamma_j^*\right] + \rho \\
&\le \frac{C_1}{k}\left\|[V]^i\right\|_1 + \rho
\end{aligned} \tag{2.8.11}$$

where the second step follows from the assumption $|\xi_i| \le \rho$, and the last step follows from Assumption **(A2)**.

Bounding $\mathbb{E}[\gamma_i]$. By (2.8.9),(2.8.10), and (2.8.11), we have

$$\mathbb{E}[\gamma_i] \le \mathbb{E}[\Sigma_{i,i}^{-1}\gamma^*] + \mathbb{E}[\Delta] \le \frac{C_1}{k}\Sigma_{i,i}^{-1} + \frac{C_1}{k}\left\|[V]^i\right\|_1 + \rho.$$

Bounding $\mathbb{E}[\gamma_i^*\gamma_i]$. First, note that

$$\begin{aligned}
\mathbb{E}[\gamma_i^*\Delta] &\le \mathbb{E}\left[\gamma_i^*\left|\sum_j V_{i,j}\gamma_j^*\right|\right] + \mathbb{E}\left[\gamma_i^*|\xi_i|\right] \\
&\le \mathbb{E}\left[\gamma_i^*\sum_j \gamma_j^*\left|V_{i,j}\right|\right] + \frac{\rho C_1}{k} \\
&= \sum_j \mathbb{E}\left[\gamma_i^*\gamma_j^*\right]\left|V_{i,j}\right| + \frac{\rho C_1}{k} \\
&= \mathbb{E}\left[(\gamma_i^*)^2\right]\left|V_{i,i}\right| + \sum_{j:j\neq i}\mathbb{E}\left[\gamma_i^*\gamma_j^*\right]\left|V_{i,j}\right| + \frac{\rho C_1}{k} \\
&\le \mathbb{E}\left[(\gamma_i^*)^2\right]\left|V_{i,i}\right| + \frac{C_1^2}{k^2}\sum_{j:j\neq i}\left|V_{i,j}\right| + \frac{\rho C_1}{k} \\
&\le \mathbb{E}\left[(\gamma_i^*)^2\right]\left|V_{i,i}\right| + \frac{C_1^2}{k^2}\left\|[V]^i\right\|_1 + \frac{\rho C_1}{k},
\end{aligned}$$

where the second and the fifth steps follow from Assumption **(A2)**. Therefore,

$$\mathbb{E}[\gamma_i^*\gamma_i] \ge \mathbb{E}\left[\gamma_i^*\left(\Sigma_{i,i}^{-1}\gamma_i^* - \alpha - \Delta\right)\right] \tag{2.8.12}$$

$$\ge \Sigma_{i,i}^{-1}\mathbb{E}\left[(\gamma_i^*)^2\right] - \frac{(\alpha+\rho)C_1}{k} - \mathbb{E}\left[(\gamma_i^*)^2\right]\left|V_{i,i}\right| - \frac{C_1^2}{k^2}\left\|[V]^i\right\|_1. \tag{2.8.13}$$

Putting together. For the first statement,

$$\begin{aligned}
\widetilde{\Sigma}_{i,i} &= 2\mathbb{E}[\gamma_i^*\gamma_i] - 2\mathbb{E}[\gamma_i^*]\mathbb{E}[\gamma_i] \\
&\geq 2\Sigma_{i,i}^{-1}\mathbb{E}\left[(\gamma_i^*)^2\right] - 2\frac{(\alpha+\rho)C_1}{k} - 2\mathbb{E}\left[(\gamma_i^*)^2\right]\left|V_{i,i}\right| - 2\frac{C_1^2}{k^2}\left\|[V]^i\right\|_1 \\
&- 2\frac{C_1^2}{k^2}\Sigma_{i,i}^{-1} - 2\frac{C_1^2}{k^2}\left\|[V]^i\right\|_1 - 2\frac{\rho C_1}{k} \\
&\geq \mathbb{E}\left[(x_i^*)^2\right]\left(2\Sigma_{i,i}^{-1} - 2\left|V_{i,i}\right|\right) - \frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}\Sigma_{i,i}^{-1} + \frac{2C_1}{k}\left\|[V]^i\right\|_1\right).
\end{aligned}$$

The second statement follows from

$$\widetilde{\Sigma}_{i,i} \leq 2\mathbb{E}[\gamma_i^*\gamma_i] \leq 2\mathbb{E}[\gamma_i^*(\Sigma_{i,i}^{-1}\gamma_i^* + \Delta)]$$

and the bound on $\mathbb{E}[\gamma_i^*\Delta]$. □

Lemma 66 (Main: Bound on $\widetilde{E}$)**.** *Suppose $|\xi_i| \leq \rho < \alpha$ for any example and every $i \in [k]$. Then for all $i, j \in [k]$ such that $i \neq j$, the following holds.*

(1) If $Z_{i,j} < 0$, then

$$\left|\widetilde{E}_{j,i}\right| \leq \frac{4C_1^2\|Z^i\|_1}{k^2(\alpha-\rho)}\left(\left|Z_{i,j}\right| + \rho\right).$$

(2) If $Z_{i,j} \geq 0$, then

$$-\frac{8C_1\rho}{k(\alpha-\rho)}\left(\frac{C_1\|Z^i\|_1}{n} + Z_{i,j}\right) - \frac{2C_1^2}{k^2}Z_{i,j} \leq \widetilde{E}_{j,i} \leq \frac{8C_1\rho}{k(\alpha-\rho)}\left(\frac{C_1\|Z^i\|_1}{n} + Z_{i,j}\right) + 2\mathbb{E}[(\gamma_j^*)^2]Z_{i,j}.$$

Proof of Lemma 66. Since $i \neq j$, we know that

$$\begin{aligned}
\widetilde{E}_{j,i} &= \mathbb{E}\left[(\gamma_j^* - (\gamma_j')^*)(\gamma_i - \gamma_i')\right] \\
&= \mathbb{E}\left[\gamma_j^*(\gamma_i - \gamma_i')\right] + \mathbb{E}\left[(\gamma_j')^*(\gamma_i' - \gamma_i)\right] \\
&= 2\mathbb{E}\left[\gamma_j^*(\gamma_i - \gamma_i')\right]
\end{aligned}$$

where the last equality follows from that $\gamma_j^*(\gamma_i - \gamma_i')$ and $(\gamma_j')^*(\gamma_i' - \gamma_i)$ has the same distribution. This quantity can be bounded by a coupling between γ_i and γ_i'. Define a new variable $\tilde{\gamma}^*$ as

$$[\tilde{\gamma}^*]_i = \begin{cases} \gamma_i^*, & \text{if } i \neq j, \\ (\gamma_j')^*, & \text{if } i = j. \end{cases}$$

By Assumption **(A2)**, conditional on γ_j^*, $\tilde{\gamma}^*$ has the same distribution as $(\gamma')^*$. Therefore, consider the variable $\tilde{\gamma}$ given by $\tilde{\gamma} = \phi_\alpha(\beta^\dagger(\beta^*\tilde{\gamma}^* + \nu'))$, we then have

$$\mathbb{E}\left[\gamma_j^*(\gamma_i - \gamma_i')\right] = \mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{\gamma}_i)\right].$$

In summary, we have

$$\widetilde{E}_{j,i} = 2\mathbb{E}[\gamma_j^*(\gamma_i - \tilde{\gamma}_i)]$$

where

$$\gamma_i = \left[\phi_\alpha\left(Z\gamma^* + \xi\right)\right]_i, \quad \xi = -\beta^\dagger NZ\gamma^* + \beta^\dagger\nu,$$
$$\tilde{x}_i = \left[\phi_\alpha\left(Z\gamma^* + \tilde{\xi}\right)\right]_i, \quad \tilde{\xi} = -\beta^\dagger NZ\tilde{\gamma}^* + \beta^\dagger\nu'.$$

Introduce the notation

$$w = Z_{i,i}\gamma_i^* + \sum_{l\neq i,j} Z_{i,l}\gamma_l^*.$$

We have

$$\gamma_i = \phi_\alpha\left(w + Z_{i,j}\gamma_j^* + \xi_i\right),$$
$$\tilde{\gamma}_i = \phi_\alpha\left(w + Z_{i,j}(\gamma_j')^* + \tilde{\xi}_i\right).$$

(1) Since $Z_{i,j} < 0$, $|\xi_i| \leq \rho$, and $|\tilde{\xi}_i| \leq \rho$, we know that when $w < \alpha - \rho$, $\gamma_i = \tilde{\gamma}_i = 0$. Then

$$\mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{\gamma}_i)\right] = \Pr[w \geq \alpha - \rho]\,\mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{\gamma}_i)|w \geq \alpha - \rho\right]. \tag{2.8.14}$$

By Property **??**, $\phi_\alpha(\cdot)$ is 1-Lipschitz, so

$$|\gamma_i - \tilde{\gamma}_i| \leq \left|Z_{i,j}\right|\left|\gamma_j^* - (\gamma_j')^*\right| + \left|\xi_i - \tilde{\xi}_i\right|,$$

which implies that

$$\begin{aligned}
\left|\mathbb{E}\left[\gamma_j^*(\gamma_i-\tilde{\gamma}_i)|w\geq\alpha-\rho\right]\right| &\leq \mathbb{E}\left[\gamma_j^*\left|Z_{i,j}\right|\left|\gamma_j^*-(\gamma_j')^*\right|+\gamma_j^*\left|\xi_i-\tilde{\xi}_i\right|\middle|w\geq\alpha-\rho\right] \\
&\leq \left|Z_{i,j}\right|\max\left\{\left|\gamma_j^*-(\gamma_j')^*\right|\right\}\mathbb{E}\left[\gamma_j^*|w\geq\alpha-\rho\right]+2\rho\mathbb{E}\left[\gamma_j^*|w\geq\alpha-\rho\right] \\
&\leq \left|Z_{i,j}\right|\max\left\{\left|\gamma_j^*-(\gamma_j')^*\right|\right\}\mathbb{E}\left[\gamma_j^*\right]+2\rho\mathbb{E}\left[\gamma_j^*\right] \\
&\leq 2\mathbb{E}\left[\gamma_j^*\right]\left(\left|Z_{i,j}\right|+\rho\right) \\
&\leq \frac{2C_1}{k}\left(\left|Z_{i,j}\right|+\rho\right). \qquad (2.8.15)
\end{aligned}$$

Now consider $\Pr[w\geq\alpha-\rho]$. Since

$$\mathbb{E}\left|w\right|\leq\left|Z_{i,i}\right|\mathbb{E}[\gamma_i^*]+\sum_{l\neq i,j}\left|Z_{i,l}\right|\mathbb{E}[\gamma_j^*]\leq\frac{C_1}{k}\|Z^i\|_1,$$

we have that

$$\Pr[w\geq\alpha-\rho]\leq\frac{\mathbb{E}\left|w\right|}{\alpha-\rho}\leq\frac{C_1\|Z^i\|_1}{k(\alpha-\rho)} \qquad (2.8.16)$$

Combining (2.8.14)(2.8.15) and (2.8.16) together completes the proof for the case when $Z_{i,j}<0$.

(2) Now consider the case when $Z_{i,j}\geq 0$. Again, we have

$$\begin{aligned}
\gamma_i &= \phi_\alpha\left(w+Z_{i,j}\gamma_j^*+\xi_i\right), \\
\tilde{\gamma}_i &= \phi_\alpha\left(w+Z_{i,j}(\gamma_j')^*+\tilde{\xi}_i\right).
\end{aligned}$$

For the analysis, introduce a variable

$$\tilde{u}_i=\phi_\alpha\left(w+Z_{i,j}\gamma_j^*+\tilde{\xi}_i\right).$$

If $(\gamma_j')^*>\gamma_j^*$, by Property **??** $\phi_\alpha(\cdot)$ is 1-Lipschitz, so

$$\tilde{\gamma}_i\leq\tilde{u}_i+Z_{i,j}\left((\gamma_j')^*-\gamma_j^*\right).$$

If $(\gamma_j')^*\leq\gamma_j^*$, by Property **??** $\phi_\alpha(\cdot)$ is non-decreasing, then

$$\tilde{\gamma}_i\leq\tilde{u}_i.$$

In any case,

$$\tilde{\gamma}_i \leq \tilde{u}_i + Z_{i,j}(\gamma'_j)^*.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{\gamma}_i)\right] &\geq \mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{u}_i)\right] - \mathbb{E}\left[\gamma_j^* Z_{i,j}(\gamma'_j)^*\right] \\ &\geq \mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{u}_i)\right] - \frac{C_1^2}{k^2} Z_{i,j}.\end{aligned}$$

So we only need to consider $\mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{u}_i)\right]$. Let G denote the event that $\gamma_i \neq 0$ or $\tilde{u}_i \neq 0$. Then by conditioning on γ_j^*, we have

$$\mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{u}_i)\right] = \mathbb{E}\left[\gamma_j^* \mathbb{E}\left[\gamma_i - \tilde{u}_i \middle| \gamma_j^*\right]\right]$$

and

$$\mathbb{E}\left[\gamma_i - \tilde{u}_i \middle| \gamma_j^*\right] = \Pr\left[G \middle| \gamma_j^*\right] \mathbb{E}\left[\gamma_i - \tilde{u}_i \middle| \gamma_j^*, G\right].$$

By Property **??** $\phi_\alpha(\cdot)$ is 1-Lipschitz, so

$$\left|\mathbb{E}\left[\gamma_i - \tilde{u}_i \middle| \gamma_j^*, G\right]\right| \leq \mathbb{E}\left[|\xi_i| + \left|\tilde{\xi}_i\right| \middle| \gamma_j^*, G\right] \leq 2\rho.$$

Now consider $\Pr\left[G \middle| \gamma_j^*\right]$. We have

$$\begin{aligned}\mathbb{E}\left[\left|w + Z_{i,j}\gamma_j^*\right| \middle| \gamma_j^*\right] &\leq \mathbb{E}\left[|w| \middle| \gamma_j^*\right] + Z_{i,j} \\ &\leq \frac{C_1}{k}\left\|Z^i\right\|_1 + Z_{i,j},\end{aligned}$$

where the first step follows from $\gamma_j^* \leq 1$ and the second step follows from the conditional independence in Assumption **(A2)**. Then by Markov's inequality,

$$\begin{aligned}\Pr\left[\gamma_i \neq 0 \middle| \gamma_j^*\right] &\leq \Pr\left[\left|w + Z_{i,j}\gamma_j^*\right| \geq \alpha - \rho \middle| \gamma_j^*\right] \\ &\leq \frac{1}{\alpha - \rho}\left(\frac{C_1}{k}\left\|Z^i\right\|_1 + Z_{i,j}\right).\end{aligned}$$

A similar argument leads to that

$$\Pr\left[\tilde{u}_i \neq 0 \middle| \gamma_j^*\right] \leq \frac{1}{\alpha - \rho}\left(\frac{C_1}{k}\left\|Z^i\right\|_1 + Z_{i,j}\right)$$

and thus

$$\Pr\left[G \middle| \gamma_j^*\right] \leq \frac{2}{\alpha - \rho}\left(\frac{C_1}{k}\left\|Z^i\right\|_1 + Z_{i,j}\right).$$

Putting things together,

$$\begin{aligned}\left|\mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{u}_i)\right]\right| &\leq \frac{4\rho}{\alpha - \rho}\left(\frac{C_1\|Z^i\|_1}{k} + Z_{i,j}\right)\mathbb{E}\left[x_j^*\right] \\ &\leq \frac{4C_1\rho}{k(\alpha - \rho)}\left(\frac{C_1\|Z^i\|_1}{k} + Z_{i,j}\right).\end{aligned}$$

This completes the proof for the lower bound.

Similarly, for the upper bound, introduce

$$u_i = \phi_\alpha\left(w + Z_{i,j}(\gamma'_j)^* + \xi_i\right).$$

Then in any case,

$$\gamma_i \leq u_i + Z_{i,j}\gamma_j^*$$

and thus

$$\mathbb{E}\left[\gamma_j^*(\gamma_i - \tilde{\gamma}_i)\right] \leq \mathbb{E}\left[\gamma_j^*(u_i - \tilde{\gamma}_i)\right] + \mathbb{E}\left[(\gamma_j^*)^2\right]Z_{i,j}.$$

The same argument as above shows that

$$\left|\mathbb{E}\left[\gamma_j^*(u_i - \tilde{\gamma}_i)\right]\right| \leq \frac{4C_1\rho}{k(\alpha - \rho)}\left(\frac{C_1\|Z^i\|_1}{k} + Z_{i,j}\right).$$

This completes the whole proof. □

Lemma 67 (Main: Bound on $\widetilde{N}$)**.** *Suppose* $\|E\|_s \leq \ell$, $\Sigma \geq (1 - \ell)\mathrm{Id}$, *and* $|\xi_j| \leq \rho < \alpha$.

*(1) If the noise is correlated (Assumption (**N1**)), then*

$$\left|\widetilde{N}_{i,j}\right| \le \frac{4C_\nu C_1}{(1-2\ell)^2 k(\alpha-\rho)} + \left|[N_s]_{i,j}\right|$$

*(2) If the noise is unbiased (Assumption (**N2**)) and $\|\beta^\dagger \nu\|_\infty \le \rho' < \alpha$, then*

$$\left|\widetilde{N}_{i,j}\right| \le \frac{2C_1 C_\nu \rho'(1+\|\beta^\dagger N\|_\infty)}{(1-2\ell)k(\alpha-\rho')} + \left|[N_s]_{i,j}\right|.$$

Proof of Lemma 67. (1) By the update rule,

$$\widetilde{N} = 2\mathbb{E}[\nu(\gamma-\gamma')^\top] + N_s.$$

Under Assumption (**N1**), we have that for every $i \in [k]$, $j \in [k]$,

$$\begin{aligned}
|\widetilde{N}_{i,j}| &= |2\mathbb{E}[\nu_i(\gamma_j - \gamma'_j)] + [N_s]_{i,j}| \\
&\le 4C_\nu \mathbb{E}[\gamma_j] + |[N_s]_{i,j}| \\
&= 4C_\nu \mathbb{E}\left[\phi_\alpha\left([Z\gamma^*]_j + \xi_j\right)\right] + |[N_s]_{i,j}|.
\end{aligned}$$

since $|\nu_i|$ is bounded by C_ν.

Now focus on the term $\mathbb{E}\left[\phi_\alpha\left([Z\gamma^*]_j + \xi_j\right)\right]$. We have

$$\left|[Z\gamma^*]_j\right| \le \|Z\|_\infty \|\gamma^*\|_\infty \le \|Z\|_\infty \le \frac{1}{1-2\ell}$$

by the fact that $\|\gamma^*\|_\infty \le 1$ in Assumption (**A2**), and the assumptions of the lemma on Σ and E. Then when $[Z\gamma^*]_j + \xi_j \ge \alpha$,

$$\phi_\alpha\left([Z\gamma^*]_j + \xi_j\right) \le [Z\gamma^*]_j + \xi_j - \alpha \le \frac{1}{1-2\ell} + \rho - \alpha \le \frac{1}{1-2\ell},$$

and thus

$$\begin{aligned}\mathbb{E}\left[\phi_\alpha\left([Z\gamma^*]_j+\xi_j\right)\right] &\le \frac{1}{1-2\ell}\Pr\left\{[Z\gamma^*]_j+\xi_j\ge\alpha\right\}\\ &\le \frac{1}{1-2\ell}\Pr\left\{\left|[Z\gamma^*]_j\right|\ge\alpha-\rho\right\}\\ &\le \frac{1}{1-2\ell}\frac{\mathbb{E}|[Z\gamma^*]_j|}{\alpha-\rho}\\ &\le \frac{1}{1-2\ell}\frac{\|Z\|_\infty}{\alpha-\rho}\,\mathbb{E}\left[\gamma_j^*\right]\\ &\le \frac{C_1}{(1-2\ell)^2k(\alpha-\rho)}\end{aligned}$$

where the last step uses the bound on $\mathbb{E}\left[\gamma_j^*\right]$ in Assumption **(A2)**. Therefore,

$$|\widetilde{N}_{i,j}|\le\frac{4C_\nu C_1}{(1-2\ell)^2k(\alpha-\rho)}+|[N_s]_{i,j}|.$$

(2) When the noise is unbiased, we have $\mathbb{E}[\nu|\gamma^*]=0$. Then $\mathbb{E}[\nu_i\gamma_j']=0$, and

$$\left|\widetilde{N}_{i,j}\right|=\left|2\mathbb{E}[\nu_i(\gamma_j-\gamma_j')]+[N_s]_{i,j}\right|\le 2\left|\mathbb{E}[\nu_i\gamma_j]\right|+\left|[N_s]_{i,j}\right|. \tag{2.8.17}$$

Consider the first term for a fixed γ^*, i.e., consider the conditional expectation $\mathbb{E}[\nu_i\gamma_j\mid\gamma^*]$. For notational simplicity, let $\widetilde{Z}=(Z-\beta^\dagger NZ)$ and $\tilde{\xi}=\beta^\dagger\nu$. Then

$$\mathbb{E}[\nu_i x_j\mid\gamma^*]=\mathbb{E}\left[\nu_i\phi_\alpha\left([Z\gamma^*]_j+\xi_j\right)\mid\gamma^*\right]=\mathbb{E}\left[\nu_i\phi_\alpha\left([\widetilde{Z}\gamma^*]_j+\tilde{\xi}_j\right)\mid\gamma^*\right].$$

We consider the following two cases about $[\widetilde{Z}\gamma^*]_j$.

(a) If $[\widetilde{Z}\gamma^*]_j\le\alpha-\rho'$, then $\phi_\alpha\left([\widetilde{Z}\gamma^*]_j+\tilde{\xi}_j\right)=0$ always holds, which implies that

$$\left|\mathbb{E}[\nu_i\gamma_j\mid\gamma^*]\right|=\mathbb{E}\left[\nu_i\phi_\alpha\left([\widetilde{Z}\gamma^*]_j+\tilde{\xi}_j\right)\mid\gamma^*\right]=0.$$

(b) If $[\widetilde{Z}\gamma^*]_j>\alpha-\rho'$, then

$$\phi_\alpha\left([\widetilde{Z}\gamma^*]_j+\tilde{\xi}_j\right)\le\phi_\alpha\left([\widetilde{Z}\gamma^*]_j+\rho'\right)\le[\widetilde{Z}\gamma^*]_j+\rho'-\alpha.$$

On the other side, by Property **??**,

$$\phi_\alpha\left([\widetilde{Z}\gamma^*]_j+\tilde{\xi}_j\right)\ge[\widetilde{Z}\gamma^*]_j+\tilde{\xi}_j-\alpha\ge[\widetilde{Z}\gamma^*]_j-\rho'-\alpha.$$

Putting together, we conclude that

$$\nu_i([\widetilde{Z}\gamma^*]_j - \alpha) - |\nu_i\rho'| \le \nu_i\phi_\alpha\left([\widetilde{Z}\gamma^*]_j + \tilde{\xi}_j\right) \le \nu_i([\widetilde{Z}\gamma^*]_j - \alpha) + |\nu_i\rho'|.$$

Note that $\mathbb{E}[\nu_i([\widetilde{Z}\gamma^*]_j - \alpha)|\gamma^*] = 0$, so

$$\left|\mathbb{E}[\nu_i\gamma_j \mid \gamma^*]\right| = \left|\mathbb{E}\left[\nu_i\phi_\alpha\left([\widetilde{Z}\gamma^*]_j + \tilde{\xi}_j\right) \mid \gamma^*\right]\right| \le \mathbb{E}[|\nu_i\rho'||\gamma^*] \le C_\nu\rho'.$$

Putting case (a) and case (b) together, we have

$$\left|\mathbb{E}[\nu_i\gamma_j \mid \gamma^*]\right| \le C_\nu\rho' \Pr\left\{[\widetilde{Z}\gamma^*]_j > \alpha - \rho'\right\} \le C_\nu\rho' \Pr\left\{\left|[\widetilde{Z}\gamma^*]_j\right| > \alpha - \rho'\right\}.$$

By definition of $\widetilde{Z}$ and the assumptions of the lemma on Σ and E,

$$\left|[\widetilde{Z}\gamma^*]_j\right| \le (1 + \|\beta^\dagger N\|_\infty)\left|[Z\gamma^*]_j\right| \le (1 + \|\beta^\dagger N\|_\infty)\,|Z|_\infty\,\gamma_j^* \le \frac{1 + \|\beta^\dagger N\|_\infty}{1 - 2\ell}\gamma_j^*. \tag{2.8.18}$$

Then

$$\Pr\left\{\left|[\widetilde{Z}\gamma^*]_j\right| > \alpha - \rho'\right\} \le \frac{\mathbb{E}\left|[\widetilde{Z}\gamma^*]_j\right|}{\alpha - \rho'} \le \frac{C_1(1 + \|\beta^\dagger N\|_\infty)}{(1 - 2\ell)k(\alpha - \rho')}.$$

The lemma then follows from (2.8.17) and (2.8.18). □

There are three terms Z, V and ξ in the above lemmas that need to be bounded. Since $Z = V + \Sigma^{-1}$, we only need to bound V and ξ in the following two lemmas, respectively.

Lemma 68 (Bound on V)**.** *Suppose* $\|E\|_s < \ell_e$ *and* $\Sigma \succeq (1 - \ell)\mathrm{Id}$. *Then*

(1) $\|V_+\|_s \le \dfrac{1 - \ell_e}{(1 - \ell)(1 - \ell_e - \ell)}\|E_-\|_s + \dfrac{\ell}{(1 - \ell)^2(1 - \ell_e - \ell)}\|E_+\|_s$,

(2) $\|V_-\|_s \le \dfrac{1 - \ell_e}{(1 - \ell)(1 - \ell_e - \ell)}\|E_+\|_s + \dfrac{\ell}{(1 - \ell)^2(1 - \ell_e - \ell)}\|E_-\|_s$,

(3) $\|V\|_s \le \dfrac{\ell_e(1 - \ell_e)}{(1 - \ell)^2(1 - \ell_e - \ell)}$,

(4) $\left|V_{i,i}\right| \le \dfrac{\ell\ell_e}{(1 - \ell)^2(1 - \ell_e - \ell)}$, $\forall i \in [k]$.

Proof of Lemma 68. Denote $T = \Sigma^{-1}\sum_{m=2}^{\infty}(-E\Sigma^{-1})^m$, so that

$$V = -\Sigma^{-1}E\Sigma^{-1} + T.$$

The following bound on $\|T\|_1$ will be useful.

$$\begin{aligned}\|T\|_1 &\le \left\|\Sigma^{-1}\right\|_1 \sum_{i=2}^{\infty} \left\|(E\Sigma^{-1})^i\right\|_1 \\ &\le \left\|\Sigma^{-1}\right\|_1 \sum_{i=2}^{\infty} \left\|E\Sigma^{-1}\right\|_1^i \\ &\le \left\|\Sigma^{-1}\right\|_1 \frac{\left\|E\Sigma^{-1}\right\|_1^2}{1-\left\|E\Sigma^{-1}\right\|_1} \\ &\le \frac{1}{(1-\ell)^3} \times \ell \times \frac{\|E\|_1}{1-\frac{\ell_e}{1-\ell}} \\ &\le \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}\|E\|_1. \end{aligned} \tag{2.8.19}$$

(1) We need to show the bound for both $\|V_+\|_1$ and $\|V_+\|_\infty$. By definition of V, for any i,

$$\|V_+\|_1 = \left\|\left[-\Sigma^{-1}E\Sigma^{-1} + T\right]_+\right\|_1.$$

Since for any β and B,

$$\|[\beta + B]_+\|_1 \le \|[\beta]_+\|_1 + \|[B]_+\|_1, \text{ and } \|[\beta]_+\|_1 \le \|\beta\|_1,$$

we have

$$\begin{aligned}\|[V_+]_i\|_1 &\le \left\|\left[-\Sigma^{-1}E\Sigma^{-1}\right]_+\right\|_1 + \|T_+\|_1 \\ &\le \frac{1}{(1-\ell)^2}\|E_-\|_1 + \|T\|_1. \end{aligned} \tag{2.8.20}$$

By (2.8.19),

$$\|T\|_1 \le \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}\|E\|_1 \le \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}(\|E_-\|_1 + \|E_+\|_1).$$

Combined with (2.8.20), it implies

$$\|[V_+]_i\|_1 \le \frac{1-\ell_e}{(1-\ell)^2(1-\ell_e-\ell)}\|E_-\|_1 + \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}\|E_+\|_1.$$

Similarly, we have

$$\left\|[V_+]^i\right\|_1 \le \frac{1-\ell_e}{(1-\ell)^2(1-\ell_e-\ell)}\|E_-\|_\infty + \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}\|E_+\|_\infty.$$

Putting things together we have

$$\|V_+\|_{\mathrm{s}} \le \frac{1-\ell_e}{(1-\ell)(1-\ell_e-\ell)}\|E_-\|_{\mathrm{s}} + \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}\|E_+\|_{\mathrm{s}}\,.$$

(2) The argument for $\|V_-\|_{\mathrm{s}}$ is similar to that for $\|V_+\|_{\mathrm{s}}$.

(3) We need to show the bound for both $\|V\|_1$ and $\|V\|_\infty$.

$$\begin{aligned}\|V\|_1 &\le \left\|-\Sigma^{-1}E\Sigma^{-1}\right\|_1 + \|T\|_1\\ &\le \frac{\ell_e}{(1-\ell)^2} + \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}\|E\|_1\\ &\le \frac{\ell_e}{(1-\ell)^2} + \frac{\ell\ell_e}{(1-\ell)^2(1-\ell_e-\ell)}\\ &= \frac{\ell_e(1-\ell_e)}{(1-\ell)^2(1-\ell_e-\ell)}\end{aligned}$$

where the second step is by (2.8.19).

Similarly, $\|V\|_\infty \le \frac{\ell_e(1-\ell_e)}{(1-\ell)^2(1-\ell_e-\ell)}$, so $\|V\|_{\mathrm{s}} \le \frac{\ell_e(1-\ell_e)}{(1-\ell)^2(1-\ell_e-\ell)}$.

(4) Now consider $V_{i,i}$. By definition of T.

$$V_{i,i} = \left[-\Sigma^{-1}E\Sigma^{-1}\right]_{i,i} + T_{i,i}.$$

Note that since $E_{i,i} = 0$, $\left[-\Sigma^{-1}E\Sigma^{-1}\right]_{i,i} = 0$. Then

$$\begin{aligned}\left|V_{i,i}\right| &= \left|T_{i,i}\right|\\ &\le \|T\|_1\\ &\le \frac{\ell}{(1-\ell)^2(1-\ell_e-\ell)}\|E\|_1\\ &\le \frac{\ell\ell_e}{(1-\ell)^2(1-\ell_e-\ell)}\end{aligned}$$

where the third step is by (2.8.19). This completes the proof. □

Lemma 69 (Bound on ξ)**.** *Suppose* $\|E\|_{\mathrm{s}} < \ell \le 1/8$ *and* $\Sigma \ge (1-\ell)\mathrm{Id}$*. Then for any* $i \in [k]$*,*

$$|\xi_i| \le \gamma := \frac{1}{1-2\ell}\left\|\beta^\dagger\right\|_\infty\|N\|_\infty + C_\nu\left\|\beta^\dagger\right\|_\infty.$$

If furthermore, $\|N\|_\infty \left\|(\beta^*)^\dagger\right\|_\infty < 1/8$*, then*

$$\left\|\beta^\dagger\right\|_\infty \le 2\left\|(\beta^*)^\dagger\right\|_\infty,$$
$$\gamma \le 3\left\|(\beta^*)^\dagger\right\|_\infty (\|N\|_\infty + C_\nu).$$

Proof of Lemma 69. First, we have

$$\|\xi\|_\infty \le \left\|\beta^\dagger N Z \gamma^*\right\|_\infty + \left\|\beta^\dagger \nu\right\|_\infty \le \left\|\beta^\dagger\right\|_\infty \|N\|_\infty \|Z\|_\infty \|\gamma^*\|_\infty + \left\|\beta^\dagger\right\|_\infty \|\nu\|_\infty.$$

Note that $\|\gamma^*\|_\infty \le 1$ and $\|\nu\|_\infty \le C_\nu$. Furthermore,

$$\|Z\|_\infty \le \frac{1}{1-2\ell}.$$

The first statement follows from combining these terms.

Now consider the second statement. We apply Lemma 70. Since

$$\begin{aligned}
\zeta &= \|E\Sigma^{-1} + (\beta^*)^\dagger N\Sigma^{-1}\|_\infty \\
&\le \|E\Sigma^{-1}\|_\infty + \|(\beta^*)^\dagger N\Sigma^{-1}\|_\infty \\
&\le \frac{1}{7} + \|(\beta^*)^\dagger\|_\infty \times \|N\|_\infty \times \|\Sigma^{-1}\|_\infty \\
&\le \frac{2}{7},
\end{aligned}$$

Lemma 70 implies that

$$\|\beta^\dagger\|_\infty \le \frac{\|\Sigma^{-1}\|_\infty}{1-\zeta}\|(\beta^*)^\dagger\|_\infty \le 2\|(\beta^*)^\dagger\|_\infty.$$

Then γ is bounded by

$$\begin{aligned}
\gamma &= \frac{1}{1-2\ell}\left\|\beta^\dagger\right\|_\infty \|N\|_\infty + C_\nu \left\|\beta^\dagger\right\|_\infty \\
&\le \frac{1}{1-2\ell} \times \left(2\|(\beta^*)^\dagger\|_\infty\right) \times \|N\|_\infty + C_\nu \times \left(2\|(\beta^*)^\dagger\|_\infty\right) \\
&\le 3\left\|(\beta^*)^\dagger\right\|_\infty (\|N\|_\infty + C_\nu).
\end{aligned}$$

□

The following is the lemma about the norm of the pseudo-inverse, which is used in Lemma 69.

Lemma 70 (Pseudo-inverse)**.** *Let* $\beta^*, N \in \mathbb{R}^{n\times k}$ *be two matrices with* $n \ge k$*. Let* $(\beta^*)^\dagger$ *be one pseudo-inverse of* β^* *such*

that $(\beta^*)^\dagger\beta^* = \mathrm{Id}$. *Let* $\beta = \beta^*(\Sigma + E) + N$ *be another matrix, with* Σ *being diagonal and*

$$\zeta := \|E\Sigma^{-1} + (\beta^*)^\dagger N\Sigma^{-1}\|_\infty.$$

satisfies $\zeta < 1$. *Then there exists a pseudo-inverse* $\beta^\dagger$ *of* β *such that* $\beta^\dagger\beta = \mathrm{Id}$ *and*

$$\|\beta^\dagger\|_\infty \le \frac{\|\Sigma^{-1}\|_\infty}{1-\zeta}\|(\beta^*)^\dagger\|_\infty.$$

Proof of Lemma 70. Consider the matrix

$$\beta^\dagger = (\Sigma + E + (\beta^*)^\dagger N)^{-1}(\beta^*)^\dagger.$$

Then by definition,

$$\begin{aligned}\beta^\dagger\beta &= (\Sigma + E + (\beta^*)^\dagger N)^{-1}(\beta^*)^\dagger\,(\beta^*(\Sigma + E) + N)\\ &= (\Sigma + E + (\beta^*)^\dagger N)^{-1}(\Sigma + E + (\beta^*)^\dagger N)\\ &= \mathrm{Id}.\end{aligned}$$

What remains is to bound $\|\beta^\dagger\|_\infty$. We have

$$\|\beta^\dagger\|_\infty \le \|(\Sigma + E + (\beta^*)^\dagger N)^{-1}\|_\infty\|(\beta^*)^\dagger\|_\infty.$$

By Taylor expansion rule, the first term on the right-hand side is

$$\begin{aligned}(\Sigma + E + (\beta^*)^\dagger N)^{-1} &= \left(\left(\mathrm{Id} + E\Sigma^{-1} + (\beta^*)^\dagger N\Sigma^{-1}\right)\Sigma\right)^{-1}\\ &= \Sigma^{-1}\left(\mathrm{Id} + E\Sigma^{-1} + (\beta^*)^\dagger N\Sigma^{-1}\right)^{-1}\\ &= \sum_{i=0}^{\infty}\Sigma^{-1}\left(-E\Sigma^{-1} - (\beta^*)^\dagger N\Sigma^{-1}\right)^i\end{aligned}$$

where we use the assumption that $\|E\Sigma^{-1} + (\beta^*)^\dagger N\Sigma^{-1}\|_\infty = \zeta < 1$. Therefore,

$$\|(\Sigma + E + (\beta^*)^\dagger N)^{-1}\|_\infty \le \|\Sigma^{-1}\|_\infty\sum_{i=0}^{\infty}\zeta^i = \frac{\|\Sigma^{-1}\|_\infty}{1-\zeta}.$$

□

2.8.6.2 Putting things together

We are now ready to prove our main theorems.

Theorem 62 (Adversarial noise, (Li et al., 2016))**.** *There exists an absolute constant $\mathcal{G}$ such that if Assumption **(A0)**-**(A3)** and **(N1)** are satisfied with $l = 1/10$, $C_2 \le 2c_2$, $C_1^3 \le \mathcal{G}c_2^2 n$, $C_\nu \le \left\{\frac{c_2^2\mathcal{G}c}{C_1^2 m}, \frac{c_2^4\mathcal{G}c}{C_1^5 k\|(\beta^*)^\dagger\|_\infty}\right\}$ for $0 \le c \le 1$, and $\left\|N^0\right\|_\infty \le \frac{c_2^2\mathcal{G}c}{C_1^3\|(\beta^*)^\dagger\|_\infty}$, then there is a choice of parameters α, η, r such that for every $0 < \epsilon, \delta < 1$ and $N = \mathrm{poly}(k, n, 1/\epsilon, 1/\delta)$, with probability at least $1 - \delta$ the following holds:*

After $T = O\left(\ln \frac{1}{\epsilon}\right)$ iterations, Algorithm 6 outputs a solution $\beta = \beta^*(\Sigma + E) + N$ where $\Sigma \ge (1-\ell)\mathrm{Id}$ is diagonal, $\|E\|_1 \le \epsilon + c/2$ is off-diagonal, and $\|N\|_1 \le c/2$.

Proof of Theorem 62. We consider the following set of parameters

$$A = \frac{c_2}{80C_1}, r = \frac{k}{c_2}, \eta = \frac{\ell}{6}.$$

Furthermore, set $\rho = B_1 \frac{c_2^2 c}{C_1^3}$ for a sufficiently small absolute constant B_1. Since $C_1 \ge k\mathbb{E}[\gamma_i^*] \ge k\mathbb{E}[(\gamma_i^*)^2] \ge c_2$, this is small enough so that

$$\rho \le \min\left\{\frac{A}{2}, \frac{c_2 A}{2048C_1}, \frac{c_2 A}{8000 \times 100C_1^2}, \frac{cc_2 A}{48000C_1^2}\right\}$$

which will be used in the proof. The proof also needs $C_1^2 \le B_1 c_2 k, C_1^3 \le B_2 c_2^2 k$ for sufficiently small absolute constants B_1 and B_2. Since $C_1 > c_2$, we only need $C_1^3 \le \mathcal{G}c_2^2 k$. Similarly, we need

$$C_\nu \le B_1 \min\left\{\frac{c(A-\rho)c_2}{nC_1}, \frac{(A-\rho)c_2}{kC_1\left\|(\beta^*)^\dagger\right\|_\infty}, \frac{(A-\rho)c_2\rho}{kC_1\left\|(\beta^*)^\dagger\right\|_\infty}, \frac{\rho}{\|(\beta^*)^\dagger\|_\infty}\right\}$$

for a sufficiently small absolute constant B_1. This can be satisfied by setting $\mathcal{G}$ small enough in the theorem assumption.

After setting the parameters needed, we now prove the theorem. We prove it by proving the following three claims by induction on t: at the beginning of iteration t,

(1) $(1-\ell)\mathrm{Id} \le \Sigma^t$,

(2) $\left\|E^t\right\|_s \le \frac{1}{8}$, and if $t > 0$

$$\left\|E_+^t\right\|_s + \beta\left\|E_-^t\right\|_s \le \left(1 - \frac{1}{25}\eta\right)\left(\left\|[E]_+^{t-1}\right\|_s + \beta\left\|[E]_-^{t-1}\right\|_s\right) + \frac{c}{10},$$

for $\beta = \frac{\sqrt{84^2+2800}-84}{2} \in (1, 8)$,

(3) $\left\|N^t\right\|_\infty \le \frac{1}{8\left\|(\beta^*)^\dagger\right\|_\infty}$, and $\|\xi^{(t)}\|_\infty \le \rho$.

Claim (1) and (2) are clearly true at $t = 0$ by the assumption on initialization. The first part of Claim (3) is true because of the assumption that $\left\|N^0\right\|_\infty \le \frac{\mathcal{G}_c}{8\mu^3\left\|(\beta^*)^\dagger\right\|_\infty}$ and that $\mu = C_1/c_2 \ge 1$. Then the second part follows from Lemma 69.

Now we assume they are true up to t, and show them for $t+1$.

(1) First consider the diagonal terms. Combining Lemma 65 and Lemma 68, we have

$$\begin{aligned}
\widetilde{\Sigma}_{i,i}^{(t)} &\ge \mathbb{E}\left[(\gamma_i^*)^2\right]\left(2(\Sigma_{i,i}^t)^{-1} - 2\left|V_{i,i}^t\right|\right) - \frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}(\Sigma_{i,i}^{(t)})^{-1} + \frac{2C_1}{k}\left\|\left[V^{(t)}\right]^i\right\|_1\right). \\
&\ge \frac{2C_2}{k}\left(0 - \frac{\ell^2}{(1-2\ell)(1-\ell)^2}\right) - \frac{2C_1}{k}\left(\alpha + \alpha + \frac{C_1}{k}\frac{1}{1-\ell} + \frac{2C_1}{k}\frac{\ell}{(1-\ell)(1-2\ell)}\right). \\
&= \frac{2C_2}{k}\left(0 - \frac{\ell^2}{(1-2\ell)(1-\ell)^2}\right) - \frac{2C_1}{k}\left(2\alpha + \frac{C_1}{k(1-\ell)(1-2\ell)}\right) \\
&> -\frac{c_2}{5k}.
\end{aligned}$$

The first inequality uses $\rho < \alpha/2$ and the last inequality is due to $\alpha \le \frac{c_2}{80C_1}$ and $C_1^2 \le \frac{c_2 k}{80}$. Therefore,

$$\Sigma_{i,i}^{t+1} = (1-\eta)\Sigma_{i,i}^t + \eta r\widetilde{\Sigma}_{i,i}^t \ge (1-\eta)\Sigma_{i,i}^t - \frac{\eta}{5}.$$

Assume for contradiction $\Sigma_{i,i}^{t+1} < 1 - \ell$. Then by the above inequality,

$$1 - \ell > \Sigma_{i,i}^{t+1} \ge (1-\eta)\Sigma_{i,i}^t - \frac{\eta}{5}.$$

which implies $\Sigma_{i,i}^t \le 1 - \ell + 2\eta$. In this case, by Lemma 65 and Lemma 68,

$$\begin{aligned}
\widetilde{\Sigma}_{i,i}^t &\ge \mathbb{E}\left[(\gamma_i^*)^2\right]\left(2(\Sigma_{i,i}^t)^{-1} - 2\left|V_{i,i}^t\right|\right) - \frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}(\Sigma_{i,i}^t)^{-1} + \frac{2C_1}{k}\left\|\left[V^t\right]^i\right\|_1\right). \\
&\ge \frac{2c_2}{k}\left(\frac{1}{1-\ell+2\eta} - \frac{\ell^2}{(1-2\ell)(1-\ell)^2}\right) - \frac{2C_1}{k}\left(2\alpha + \frac{C_1}{k(1-\ell)(1-2\ell)}\right) \\
&> \frac{c_2}{k}.
\end{aligned}$$

Then

$$\Sigma_{i,i}^{t+1} = (1-\eta)\Sigma_{i,i}^t + \eta r\widetilde{\Sigma}_{i,i}^t = (1-\eta)\Sigma_{i,i}^t + \eta > \Sigma_{i,i}^t,$$

which is a contradiction. Therefore, $(1-\ell)\mathrm{Id} \preceq \Sigma^t$.

(2) Now consider the off-diagonal terms. We shall split them into the positive part and the negative part. By the

update rule, for any $i \in [k]$,

$$\left\|\left[E_+^{t+1}\right]_i\right\|_1 \le (1-\eta)\left\|\left[E_+^t\right]_i\right\|_1 + \eta r\left\|\left[\widetilde{E}_+^t\right]_i\right\|_1.$$

Recall the notations

$$Z^t = (\Sigma^t + E^t)^{-1} = (\Sigma^t)^{-1} + V^t,$$
$$V^t = (\Sigma^t)^{-1}\sum_{i=1}^{\infty}(-E^t(\Sigma^t)^{-1})^i$$

By Lemma 66, we have

$$\begin{aligned}\left\|\left[\widetilde{E}_+^t\right]_i\right\|_1 \le & \underbrace{\sum_{j\neq i}\frac{4C_1^2}{k^2(\alpha-\rho)}\left\|\left[Z^t\right]^i\right\|_1\left(\left|\left[Z_-^t\right]_{i,j}\right|+\rho\right)}_{T1} \\ & + \underbrace{\sum_{j\neq i}\frac{8C_1\rho}{k(\alpha-\rho)}\left(\frac{C_1}{k}\left\|\left[Z^t\right]^i\right\|_1 + \left|\left[Z_+^t\right]_{i,j}\right|\right)}_{T2} \\ & + \underbrace{\sum_{j\neq i}2\mathbb{E}[(\gamma_j^*)^2]\left|\left[Z_+^t\right]_{i,j}\right|}_{T3}.\end{aligned}$$

First, by Lemma 68,

$$\left\|\left[Z^t\right]^i\right\|_1 \le \left[\left(\Sigma^t\right)^{-1}\right]_{i,i} + \left\|\left[V^t\right]^i\right\|_1 \le \frac{1}{1-2\ell}$$

Now consider Z_+^t and Z_-^t. We have

$$\sum_{j:j\neq i}\left|\left[Z_-^t\right]_{i,j}\right| \le \left\|\left[V_-^t\right]_i\right\|_1, \quad \sum_{j:j\neq i}\left|\left[Z_+^t\right]_{i,j}\right| \le \left\|\left[V_+^t\right]_i\right\|_1.$$

Therefore,

$$\begin{aligned}T1 &\le \frac{8C_1^2}{k^2(\alpha-\rho)}\left\|\left[V_-^t\right]_i\right\|_1 + \frac{8C_1^2\rho}{k(\alpha-\rho)}, \\ T2 &\le \frac{16C_1^2\rho}{k(\alpha-\rho)} + \frac{8C_1\rho}{k(\alpha-\rho)}\left\|\left[V_+^t\right]_i\right\|_1, \\ T3 &\le \frac{2C_2}{n}\left\|\left[V_+^t\right]_i\right\|_1.\end{aligned}$$

and thus we have

$$\left\|\left[\widetilde{E}_+^t\right]_i\right\|_1 \le \frac{8C_1^2}{k^2(\alpha-\rho)}\left\|\left[V_-^t\right]_i\right\|_1 + \left(\frac{2C_2}{n} + \frac{8C_1\rho}{k(\alpha-\rho)}\right)\left\|\left[V_+^t\right]_i\right\|_1 + \frac{24C_1^2\rho}{k(\alpha-\rho)}.$$

Similarly, for any $i \in [k]$,

$$\left\|\left[\widetilde{E}_+^t\right]^i\right\|_1 \le \frac{8C_1^2}{k^2(\alpha-\rho)}\left\|\left[V_-^t\right]^i\right\|_1 + \left(\frac{2C_2}{n} + \frac{8C_1\rho}{k(\alpha-\rho)}\right)\left\|\left[V_+^t\right]^i\right\|_1 + \frac{24C_1^2\rho}{k(\alpha-\rho)}.$$

Putting the two together, we have

$$\left\|\widetilde{E}_+^t\right\|_{\mathrm{s}} \le \frac{8C_1^2}{k^2(\alpha-\rho)}\left\|V_-^t\right\|_{\mathrm{s}} + \left(\frac{2C_2}{n} + \frac{8C_1\rho}{k(\alpha-\rho)}\right)\left\|V_+^t\right\|_{\mathrm{s}} + \frac{24C_1^2\rho}{k(\alpha-\rho)}. \tag{2.8.21}$$

By Lemma 68 and $\ell \le \frac{1}{8}$, we have:

$$\left\|V_+^t\right\|_{\mathrm{s}} \le \frac{32}{21}\left\|E_-^t\right\|_{\mathrm{s}} + \frac{32}{147}\left\|E_+^t\right\|_{\mathrm{s}},$$
$$\left\|V_-^t\right\|_{\mathrm{s}} \le \frac{32}{21}\left\|E_+^t\right\|_{\mathrm{s}} + \frac{32}{147}\left\|E_-^t\right\|_{\mathrm{s}}$$

So (2.8.21) becomes

$$\left\|\widetilde{E}_+^t\right\|_{\mathrm{s}} \le \left(\frac{64C_2}{147k} + \frac{256C_1\rho}{147k(\alpha-\rho)} + \frac{256C_1^2}{21k^2(\alpha-\rho)}\right)\left\|E_+^t\right\|_{\mathrm{s}} \tag{2.8.22}$$
$$+\left(\frac{64C_2}{21k} + \frac{256C_1\rho}{21k(\alpha-\rho)} + \frac{256C_1^2}{147n^2(\alpha-\rho)}\right)\left\|E_-^t\right\|_{\mathrm{s}} + \frac{24C_1^2\rho}{k(\alpha-\rho)}. \tag{2.8.23}$$

Now consider the negative part. The same argument as above leads to

$$\left\|\widetilde{E}_-^t\right\|_{\mathrm{s}} \le \left(\frac{64C_1^2}{147k^2} + \frac{256C_1\rho}{147k(\alpha-\rho)} + \frac{256C_1^2}{21n^2(\alpha-\rho)}\right)\left\|E_+^t\right\|_{\mathrm{s}}$$
$$+\left(\frac{64C_1^2}{21k^2} + \frac{256C_1\rho}{21k(\alpha-\rho)} + \frac{256C_1^2}{147k^2(\alpha-\rho)}\right)\left\|E_-^t\right\|_{\mathrm{s}} + \frac{24C_1^2\rho}{k(\alpha-\rho)}. \tag{2.8.24}$$

Note the difference between (2.8.23) and (2.8.24): $\frac{C_2}{k}$ in the former is replaced by $\frac{C_1^2}{k^2}$ in the latter, which is much smaller. This is crucial for our proof, which will be clear below.

For simplicity, we introduce the following notations:

$$a_t := \left\|E_+^t\right\|_{\mathrm{s}}, \quad b_t := \left\|E_-^t\right\|_{\mathrm{s}}.$$

Then by the update rule, we have

$$a_{t+1} \le (1-\eta)a_t + \eta r \left\|\widetilde{E}_+^t\right\|_{\mathrm{s}},$$
$$b_{t+1} \le (1-\eta)b_t + \eta r \left\|\widetilde{E}_-^t\right\|_{\mathrm{s}}.$$

Plugging in (2.8.23)and since $r = \frac{k}{c_2} \le \frac{2k}{C_2}$, we have

$$\begin{aligned}
a_{t+1} &\le (1-\eta)a_t + \eta\frac{2k}{C_2}\left(\frac{64C_2}{147k} + \frac{256C_1\rho}{147k(\alpha-\rho)} + \frac{256C_1^2}{21k^2(\alpha-\rho)}\right)a_t \\
&\quad + \eta\frac{2k}{C_2}\left(\frac{64C_2}{21k} + \frac{256C_1\rho}{21k(\alpha-\rho)} + \frac{256C_1^2}{147k^2(\alpha-\rho)}\right)b_t + \eta\frac{2n}{C_2}\frac{24C_1^2\rho}{k(\alpha-\rho)} \\
b_{t+1} &\le (1-\eta)b_t + \eta\frac{2n}{C_2}\left(\frac{64C_1^2}{147k^2} + \frac{256C_1\rho}{147k(\alpha-\rho)} + \frac{256C_1^2}{21k^2(\alpha-\rho)}\right)a_t \\
&\quad + \eta\frac{2k}{C_2}\left(\frac{64C_1^2}{21k^2} + \frac{256C_1\rho}{21k(\alpha-\rho)} + \frac{256C_1^2}{147k^2(\alpha-\rho)}\right)b_t + \eta\frac{2k}{C_2}\frac{24C_1^2\rho}{k(\alpha-\rho)}.
\end{aligned}$$

When $\frac{512C_1\rho}{C_2(\alpha-\rho)} \le \frac{1}{2}$ and $\frac{512C_1^2}{C_2k(\alpha-\rho)} \le \frac{1}{14}$,

$$\begin{aligned}
a_{t+1} &\le (1-\eta)a_t + \frac{129}{147}\eta a_t + \frac{129}{21}\eta b_t + \eta\frac{48C_1^2\rho}{C_2(\alpha-\rho)} \\
&\le \left(1 - \frac{18}{147}\eta\right)a_t + \frac{129}{21}\eta b_t + \eta\frac{48C_1^2\rho}{C_2(\alpha-\rho)}
\end{aligned}$$

Similarly, when $\frac{512C_1\rho}{C_2(\alpha-\rho)} \le \frac{1}{2}$ and $\frac{512C_1^2}{C_2k(\alpha-\rho)} \le \frac{1}{14}$, and furthermore, $\frac{128C_1^2}{C_2k} \le \frac{1}{4}$,

$$\begin{aligned}
b_{t+1} &\le (1-\eta)b_t + \frac{1}{100}\eta a_t + \frac{1}{25}\eta b_t + \eta\frac{48C_1^2\rho}{C_2(\alpha-\rho)} \\
&\le \left(1 - \frac{24}{25}\eta\right)b_t + \frac{1}{100}\eta a_t + \eta\frac{48C_1^2\rho}{C_2(\alpha-\rho)}
\end{aligned}$$

Let $h = \frac{48C_1^2\rho}{C_2(\alpha-\rho)}$, we then have:

$$\begin{aligned}
a_{t+1} &\le \left(1 - \frac{3}{25}\eta\right)a_t + 7\eta b_t + \eta h, \\
b_{t+1} &\le \left(1 - \frac{24}{25}\eta\right)b_t + \frac{1}{100}\eta a_t + \eta h.
\end{aligned}$$

Now set $\beta = \frac{\sqrt{84^2+2800}-84}{2}$, so that

$$\begin{aligned} a_{t+1} + \beta b_{t+1} &\le \left(1 - \frac{3}{25}\eta\right) a_t + 7\eta b_t + \eta h + \left(\beta - \frac{24}{25}\eta\beta\right) b_t + \frac{\beta}{100}\eta a_t + \eta\beta h \\ &= \left(1 - \frac{3}{25}\eta + \frac{\beta}{100}\eta\right)(a_t + \beta b_t) + \eta(1+\beta)h \\ &\le \left(1 - \frac{1}{25}\eta\right)(a_t + \beta b_t) + 9\eta h, \end{aligned}$$

where the last inequality follows from that $\beta < 8$.

Note that the recurrence is true up to $t+1$. Using Lemma 81 to solve this recurrence, we obtain

$$a_t + b_t \le a_0 + b_0 + 250h \le \frac{1}{10} + 250h \le \frac{1}{8}$$

when $\frac{4000C_1^2\rho}{C_2(\alpha-\rho)} \le \frac{1}{100}$. Moreover, we know that

$$\left\|E^{t+1}\right\|_{\mathrm{s}} \le a_{t+1} + \beta b_{t+1} \le \left(1 - \frac{1}{25}\eta\right)^t + 250h.$$

(3) Finally, consider the noise term. Set the sample size D to be large enough, so that by Lemma 67, we have

$$\begin{aligned} \left|\widetilde{N}_{i,j}^t\right| &\le \frac{4C_\nu C_1}{(1-2\times\ell)^2 k(\alpha-\rho)} + \left|[N_s^t]_{i,j}\right| \\ &\le \frac{8C_\nu C_1}{k(\alpha-\rho)}. \end{aligned}$$

Then by the update rule, we have $\left|N_{i,j}^{t+1}\right| \le \frac{8C_\nu C_1}{(\alpha-\rho)c_2}$. Then

$$\left\|N^{t+1}\right\|_\infty \le k \max_{i,j}\left|N_{i,j}^{t+1}\right| \le \frac{8kC_\nu C_1}{(\alpha-\rho)c_2} \le \frac{1}{8\left\|(\beta^*)^\dagger\right\|_\infty}$$

where the last inequality is due to

$$C_\nu \le \frac{(\alpha-\rho)c_2}{64nC_1\left\|(\beta^*)^\dagger\right\|_\infty}.$$

On the other hand, by Lemma 69, we have

$$\begin{aligned} \|\xi^{t+1}\|_\infty &\le 3\|(\beta^*)^\dagger\|_\infty(\|N^{t+1}\|_\infty + C_\nu) \\ &\le 3\|(\beta^*)^\dagger\|_\infty\left(\frac{8kC_\nu C_1}{(\alpha-\rho)c_2} + C_\nu\right) \le \rho \end{aligned}$$

where the last inequality is due to

$$C_\nu \le \frac{(\alpha-\rho)c_2\rho}{48kC_1\|(\beta^*)^\dagger\|_\infty}, \text{ and } C_\nu \le \frac{\rho}{6\|(\beta^*)^\dagger\|_\infty}.$$

We also have (which will be useful in proving the final bound)

$$\left\|N^{t+1}\right\|_1 \le n \max_{i,j} \left|N_{i,j}^{t+1}\right| \le \frac{8nC_\nu C_1}{(\alpha-\rho)c_2} \le \frac{c}{10}$$

where the last inequality is due to

$$C_\nu \le \frac{c(\alpha-\rho)c_2}{80nC_1}.$$

Now, we shall prove the theorem statements. Recall that solving the recurrence about a_t and b_t leads to

$$\left\|E^{t+1}\right\|_{\mathrm{s}} \le a_{t+1} + \beta b_{t+1} \le \left(1 - \frac{1}{25}\eta\right)^t + 250h.$$

Since the setting of ρ makes sure $h = O(c)$, when $t = O\left(\ln \frac{1}{\epsilon}\right)$, we have the second statement $\left\|\widehat{E}\right\|_{\mathrm{s}} \le \epsilon + \frac{c}{2}$. Note that

$$\beta^*\widehat{\Sigma} = \beta - \beta^*\widehat{E} - \widehat{N}$$

and

$$\left\|\left[\beta^*\widehat{\Sigma}\right]_i\right\| = \widehat{\Sigma}_{i,i}, \quad \|\beta\|_1 = 1, \quad \left\|\beta^*\widehat{E}\right\|_1 = \left\|\widehat{E}\right\|_1,$$

so we have

$$\begin{aligned}\widehat{\Sigma}_{i,i} &\ge \|\beta\|_1 - \left\|\widehat{E}\right\|_1 - \left\|\widehat{N}\right\|_1 \\ &\ge 1 - \epsilon - c.\end{aligned}$$

Similarly,

$$\begin{aligned}\widehat{\Sigma}_{i,i} &\le \|\beta\|_1 + \left\|\widehat{E}\right\|_1 + \left\|\widehat{N}\right\|_1 \\ &\le 1 + \epsilon + c.\end{aligned}$$

Then the final statement of the theorem follows by replacing c with $c/4$. This completes the proof. □

Theorem 63 (Unbiased noise, (Li et al., 2016)). *If Assumption **(A0)**-**(A3)** and **(N2)** are satisfied with $C_\nu = \frac{c_2\mathcal{G}\sqrt{ck}}{C_1 \max\{n,k\}\|(\beta^*)^\dagger\|_\infty}$ and the other parameters set as in Theorem 62, then the same guarantee holds.*

Proof. The proof is similar to that of Theorem 62, except using the second bound for unbiased noise in Lemma 67. We highlight the different part, that is, the induction on the noise term.

In the induction, by Lemma 67 we have when N is large enough,

$$\left|\widetilde{N}^t_{i,j}\right| \le \frac{2C_1C_\nu\rho'(1+\|\beta^\dagger N^t\|_\infty)}{(1-2\ell)k(\alpha-\rho')} + \left|\left[N^t_s\right]_{i,j}\right| \le \frac{3C_1C_\nu\rho'(1+\|\beta^\dagger N^t\|_\infty)}{k(\alpha-\rho')}.$$

By Lemma 69 and the induction, we have $\|\beta^\dagger N^t\|_\infty \le 1/4$. Furthermore, $\rho' \le C_\nu\|\beta^\dagger\|_\infty \le 2C_\nu\|(\beta^*)^\dagger\|_\infty$ and the parameter setting makes sure $\rho' \le \alpha/2$. Then

$$\left|\widetilde{N}^t_{i,j}\right| \le \frac{16C_\nu^2C_1\left\|(\beta^*)^\dagger\right\|_\infty}{k\alpha}.$$

Then by the update rule, we have

$$\left|N^{t+1}_{i,j}\right| \le \frac{32C_\nu^2C_1\left\|(\beta^*)^\dagger\right\|_\infty}{c_2\alpha}$$

and

$$\left\|N^{t+1}\right\|_\infty \le \frac{32nC_\nu^2C_1\left\|(\beta^*)^\dagger\right\|_\infty}{c_2\alpha} \le \frac{1}{8\left\|(\beta^*)^\dagger\right\|_\infty} \tag{2.8.25}$$

by the definition of α, and $C_\nu \le \frac{1}{256}\frac{c_2}{C_1}\frac{\sqrt{k}}{k\|(\beta^*)^\dagger\|_\infty}$. This completes the induction for the noise.

Also, in proving the final bounds, we have

$$\left\|N^{t+1}\right\|_1 \le \frac{32nC_\nu^2C_1\left\|(\beta^*)^\dagger\right\|_\infty}{c_2\alpha} \le \frac{c}{10} \tag{2.8.26}$$

by the definition of α, and

$$C_\nu \le \frac{c}{320}\frac{c_2}{C_1}\frac{\sqrt{k}}{\max\left\{n, k\left\|(\beta^*)^\dagger\right\|_\infty\right\}} \le \frac{\sqrt{c}}{320}\frac{c_2}{C_1}\frac{1}{\sqrt{n\left\|(\beta^*)^\dagger\right\|_\infty}}$$

where the last inequality can be shown by consider the two cases when $\left\|(\beta^*)^\dagger\right\|_\infty \le n/k$ and $\left\|(\beta^*)^\dagger\right\|_\infty \ge n/k$. The rest of the proof is the same as in Theorem 62. □

2.8.7 Results for general proportions: Equilibration

Algorithm 7 ColumnUpdate

Input: A matrix β, a threshold value α, a step size η, ratios $\{r_j : j \in [k]\}$, iteration number T, a subset $S \subseteq [k]$, sample size D

1: Set $\beta^{(0)} = \beta$
2: **for** $t = 0 \to T - 1$ **do**
3:

$$\forall i \in S, [\beta^{t+1}]_i = \left[(1-\eta)\beta^t + r_i \eta \tilde{\mathbb{E}}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right]\right]_i \tag{2.8.27}$$

4: **end for**

Output: $\hat{\beta} = \beta^{(T)}$

Algorithm 8 Rescale

Input: A matrix β, a threshold value α, a step size η, ratios $\{r_j : j \in [k]\}$, iteration number T, and a set $S \subseteq [k]$, $\epsilon \in (0, 1)$.

1: Let $\tilde{\beta} = \text{ColumnUpdate}(\beta, \alpha, \eta, \{r_j\}_j, T, S, D)$
2: **for** $i \in S$ **do**
3: Set $[\hat{\beta}]_i = \frac{1}{1-\epsilon}[\tilde{\beta}]_i$
4: **end for**

Output: $\hat{\beta}$

Algorithm 9 Equilibration

Input: β, α, η, T, and $\epsilon \in (0, 1)$, λ, D

1: $S \leftarrow \emptyset$, $D \leftarrow \text{Id}$
2: **while** $|S| \le k$ **do**
3: $M_j \leftarrow \hat{\mathbb{E}}[\gamma_j^2]$ for $j \notin S$ using D examples
4: **while** $\max_{j \notin S} M_j < \lambda$ **do**
5: $\beta \leftarrow \text{Rescale}(\beta, \alpha, \eta, \{3/(5M_j) : j \in [k]\}, T, S, \epsilon, D)$
6: $\lambda \leftarrow (1-\epsilon)\lambda$, $D_{j,j} \leftarrow D_{j,j}/(1-\epsilon)$
7: $M_j \leftarrow (1-\epsilon)^2 M_j$ for $j \in S$, and $M_j \leftarrow \hat{\mathbb{E}}[\gamma_j^2]$ for $j \notin S$ using D examples
8: **end while**
9: $S \leftarrow S \cup \{j : M_j \ge \lambda\}$
10: **end while**

Output: β

When the topics have various proportions (i.e., $\mathbb{E}[(\gamma_i^*)^2]$ varies for different i), we propose Algorithm 9 for balancing them. Recall the idea: instead of solving $\tilde{f} \approx \beta^* \gamma^*$, we could also solve $\tilde{f} \approx [\beta^* D][(D)^{-1}\gamma^*]$ for a positive diagonal matrix D. Our goal is to find $\beta = \beta^* D(\Sigma + E) + N$ so that Σ is large, E, N are small, while $\mathbb{E}[(\gamma_i^*)^2]/D_{i,i}^2$ is with in a factor of 2 from each other.

The algorithm works at stages and keeps a working set S of column index i such that $\mathbb{E}[(\gamma_i^*)^2]/D_{i,i}^2$ is above a threshold λ. At each stage, it only updates the columns in S; at the end of the stage, it increases these columns by a small factor so that $\mathbb{E}[(\gamma_i^*)^2]/D_{i,i}^2$ decreases. Then it decreases the threshold λ, and add more columns to the working

set and repeat. In this way, $\mathbb{E}[(\gamma_i^*)^2]/D_{i,i}^2 (i \in S)$ are always balanced; in particular, they are balanced at the end when $S = [n]$. Formally,

Theorem 71 (Main: Equilibration)**.** *If there exists an absolute constant $\mathcal{G}$ such that Assumption (**A1**)-(**A3**) and (**N1**) are satisfied with $l = 1/50$, $C_1^3 \le \mathcal{G} c_2^2 k$, $\max\left\{C_\nu, \|N^{(0)}\|_\infty\right\} \le \frac{\mathcal{G} c_2^4}{C_1^5 k \|(\beta^*)^\dagger\|_\infty}$, and additionally $\Sigma^{(0)} \preceq (1-\ell)\mathrm{Id}$, and $E \ge 0$ entry-wise, then there exist α, η, T, λ such that for sufficiently small $\epsilon > 0$ and sufficiently large $D = \mathrm{poly}(k, n, 1/\epsilon, 1/\delta)$ the following hold with probability at least $1-\delta$: Algorithm 9 outputs a solution $\beta = \beta^* D(\Sigma + E) + N$ where $\Sigma \ge (1-\ell)\mathrm{Id}$ is diagonal, $\|E\|_s \le \gamma\ell$ is off-diagonal, $\|N\|_\infty \le 2\|N^{(0)}\|_\infty$, and D is diagonal and satisfies*

$$\frac{\max_{i\in[k]} \frac{1}{D_{i,i}^2}\mathbb{E}[(\gamma_i^*)^2]}{\min_{j\in[k]} \frac{1}{D_{j,j}^2}\mathbb{E}[(\gamma_j^*)^2]} \le 2.$$

*If Assumption (**A1**)-(**A3**) and (**N2**) are satisfied with the same parameters except $\max\left\{C_\nu, \|N^{(0)}\|_\infty\right\} \le \min\left\{\sqrt{\frac{\mathcal{G} c_2^4}{C_1^5 k}} \frac{1}{\|(\beta^*)^\dagger\|_\infty}, \frac{\mathcal{G} c_2^2}{C_1^3 \|(\beta^*)^\dagger\|_\infty}\right\}$, then the same guarantees hold.*

Now, we can view $\beta^* D$ as the ground-truth feature matrix and $D^{-1}\gamma^*$ as the weights. Then applying Algorithm 6 with β can recover $\beta^* D$, and after normalization we get β^*.

The initialization condition of the theorem can be achieved by the popular practical heuristic that sets the columns of $\beta^{(0)}$ to reasonable almost pure data points. It is generally believed that it gives $E_{i,j}^{(0)} \ge 0$ and $N^{(0)} = 0$. We note that the parameters are not optimized; the algorithm can potentially tolerate much better initialization.

Intuition. Before delving into the specifics of the algorithm, it will be useful to provide a high-level outline of the proof. As described above, the algorithm makes use of the fact that samples from a ground truth matrix β^* and distribution γ^* can equivalently be viewed as coming from the ground truth matrix $\beta^* D$ and distribution $D^{-1}x^*$, for some diagonal matrix D. Therefore, the goal is to find a D such that the features are balanced:

$$\frac{\max_{i\in[k]} \frac{\mathbb{E}[(\gamma_i^*)^2]}{D_{i,i}^2}}{\min_{i\in[n]} \frac{\mathbb{E}[(\gamma_i^*)^2]}{D_{i,i}^2}} \le \kappa.$$

The algorithm will implicitly calculate such a D gradually. Namely, at any point in time, the algorithm will have an active set $S \subseteq [k]$ of features, which are balanced, i.e.

$$\frac{\max_{i\in[k]} \frac{\mathbb{E}[(\gamma_i^*)^2]}{D_{i,i}^2}}{\min_{i\in S} \frac{\mathbb{E}[(\gamma_i^*)^2]}{D_{i,i}^2}} \le \kappa. \tag{2.8.28}$$

It is clear that when $S = [k]$ the algorithm achieves the goal. Our algorithm begins with $S = \emptyset$ and gradually increase S until $S = [k]$.

The mechanism for increasing S will be as follows. Given S, β is of the form

$$\beta = \beta^* D(\Sigma + E) + N$$

with

$$E = \begin{bmatrix} E_{1,1} & E_{1,2} \\ E_{2,1} & E_{2,2} \end{bmatrix}$$

where the columns of β are sorted such that the first $|S|$ columns correspond to the features of S, and $E_{1,1} \in \mathbb{R}^{|S|\times|S|}$, $E_{2,1} \in \mathbb{R}^{(k-|S|)\times|S|}$, $E_{1,2} \in \mathbb{R}^{|S|\times(k-|S|)}$, $E_{2,2} \in \mathbb{R}^{(k-|S|)\times(k-|S|)}$. Then scaling up the columns of β indexed by S by a factor of $\frac{1}{1-\epsilon}$ is equivalent to

(1) scaling up the columns of D indexed by S by a factor of $\frac{1}{1-\epsilon}$ and

(2) scaling up the columns of $E_{2,1}$ by a factor of $\frac{1}{1-\epsilon}$ and

(3) scaling down the columns of $E_{1,2}$ by a factor of $1 - \epsilon$.

Therefore, to increase the set S, the algorithm will scale up the columns of β indexed by S, until some $j \notin S$ satisfies

$$\max_{i\in[k]} \frac{\mathbb{E}[(\gamma_i^*)^2]}{D_{i,i}^2} \le \kappa \frac{\mathbb{E}[(\gamma_j^*)^2]}{D_{j,j}^2}.$$

Then it can add j into S while keeping the corresponding features balanced as in (2.8.28). Note that we do not need to explicitly maintain D, though it can be calculated along with the scaling. Further note that the values of $\mathbb{E}[(\gamma_i^*)^2]$ are not known but they can be estimated using the current β.

However, there is still one caveat: E should be kept small, so that at the end of the algorithm, we still have a good initialization β. For this reason, the algorithm additionally maintains that for a small constant $1 < \gamma < 2$,

$$\begin{aligned} \left\|E_{1,1}\right\|_{\mathrm{s}} \le \gamma\ell, \quad \left\|E_{1,2}\right\|_{\mathrm{s}} \le \ell, \\ \left\|E_{2,1}\right\|_{\mathrm{s}} \le \gamma\ell, \quad \left\|E_{2,2}\right\|_{\mathrm{s}} \le \ell. \end{aligned} \tag{2.8.29}$$

Since scaling up β will scale up $E_{2,1}$, we will need to first decrease $\left\|E_{2,1}\right\|_{\mathrm{s}}$ before the scaling step. The key observation is that by applying our training algorithm only on the columns indexed by S, $\left\|E_{1,1}\right\|_{\mathrm{s}}$ and $\left\|E_{2,1}\right\|_{\mathrm{s}}$ will be decreased, while $\left\|E_{1,2}\right\|_{\mathrm{s}}$ and $\left\|E_{2,2}\right\|_{\mathrm{s}}$ unchanged. On a high level, using the fact that the matrix $E_{1,2}$ has no negative entries (which

we get by virtue of our initialization), and the fact that the contribution in the updates to the entry $(E_{1,1})_{i,j}$ mostly comes from $(E_{1,1})_{j,i}$ (i.e. the matrix $E_{1,1}$ in the first order contribution "updates itself"), and the fact that the features in S are balanced, we can show that after sufficiently many updates, the symmetric norm of $E_{1,1}$ and $E_{2,1}$ drops by a reasonable amount: $\left\|E_{1,1}\right\|_{\mathrm{s}} \leq (\gamma-1)\ell$ and $\left\|E_{2,1}\right\|_{\mathrm{s}} \leq (1-\epsilon)(\gamma-1)\ell$. Now, we can do the scaling step without hurting the invariant 2.8.29.

Organization. The result of the section is as follows. We first prove in Section 2.8.7.1 that applying our training algorithm only on the columns indexed by S will decrease $\left\|E_{1,1}\right\|_{\mathrm{s}}$ and $\left\|E_{2,1}\right\|_{\mathrm{s}}$. Then in Section 2.8.7.2 we analyze the scaling step, and show that the invariant (2.8.29) is maintained. In Section 2.8.7.3, we show how to increase S while maintaining the invariant (2.8.28), where the main technical details are about how to estimate $\mathbb{E}[(\gamma_i^*)^2]$.

2.8.7.1 Equilibration: ColumnUpdate

In this subsection, we focus on the update step, bounding the changes of Σ, E, and N.

First recall some notations. Let $\beta = \beta^*(\Sigma + E) + N$ where Σ is diagonal, E is off diagonal, and N is the component outside the span of β^*.[5] Given the set $S \subseteq [k]$ and a matrix $M \in \mathbb{R}^{k\times k}$, let $M_{1,1}$ denote the submatrix indexed by $S \times S$, and $M_{2,1}$ denote the submatrix indexed by $([k]-S)\times S$, $M_{1,2}$ denote the submatrix indexed by $S \times ([k]-S)$, and $M_{2,2}$ denote the submatrix indexed by $([k]-S)\times([k]-S)$. [6] In the special case when $S = [s]$ where $s = |S|$,

$$M = \begin{bmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2} \end{bmatrix}.$$

Also, let M_S denote the submatrix formed by the columns indexed by S, and M_{-S} the submatrix formed by the other columns. [7]

The input β^0 of Algorithm 7 can be written as $\beta^0 = \beta^*(\Sigma^0 + E^0) + N^0$ where Σ^0 is diagonal, and E^0 is off diagonal. Define $E^0_{1,1}, E^0_{1,2}, E^0_{2,1}$ and $\hat{E}_{2,2}$ as described above. Similarly, define $\hat{E}_{1,1}, \hat{E}_{1,2}, \hat{E}_{2,1}$ and $\hat{E}_{2,2}$ for the output $\hat{\beta} = \beta^*(\hat{\Sigma} + \hat{E}) + \hat{N}$ of Algorithm 7. Finally, define N^0_S, N^0_{-S}, $\hat{N}_S$, and $\hat{N}_{-S}$ as described above.

The main result of the subsection is Lemma 72.

[5] Note that β^* here can be any ground-truth matrix; in particular, later Lemma 72 will be applied where β^* in the lemma corresponds to $\beta^* D$ in the intuition described above.

[6] These notations will be used for $M = E$, $M = \widetilde{E}$, and related matrices.

[7] These notations will be used for $M = N$ or $M = \widetilde{N}$, and related matrices.

Lemma 72 (Main: ColumnUpdate)**.** *Define*

$$R_j = \mathbb{E}[(\gamma_j^*)^2], \quad R = \max_{j\in[k]} R_j, \quad r = \max_{j\in S} r_j, \tag{2.8.30}$$

$$h_1 = r\frac{8C_1(C_1+1)\rho}{(1-\ell-\beta\ell)k(\alpha-\rho)} + \frac{4C_1^2}{(1-\ell-\beta\ell)k^2(\alpha-\rho)} r\left(\frac{1}{(1-\ell-\beta\ell)}+1\right), \tag{2.8.31}$$

$$h_2 = r\frac{R\beta^2\ell^2}{(1-\ell)^2(1-\ell-\beta\ell)} + \frac{12C_1(C_1+1)}{k^2(\alpha-\rho)(1-\ell-\beta\ell)}\left(\frac{1}{1-\ell-\beta\ell}+k\rho\right) r, \tag{2.8.32}$$

$$h = h_1 + h_2, \tag{2.8.33}$$

$$U_a = \frac{8rC_\nu C_1}{\alpha-\rho}, \tag{2.8.34}$$

$$U_n = \frac{10rC_1C_\nu^2\left\|(\beta^*)^\dagger\right\|_\infty}{(1-2\ell)(\alpha-2C_\nu\left\|(\beta^*)^\dagger\right\|_\infty)}. \tag{2.8.35}$$

Suppose $\ell \le 1/8$*,* B *is a constant with* $B\ell \le 1/2$*,* $\gamma \in (1,2)$*,* $\epsilon \in (0,1)$*. The initialization satisfies* $(1-\ell)\mathrm{Id} \le \Sigma^0$*,* $\left\|E_{1,1}^0\right\|_{\mathrm{s}} \le \gamma\ell, \left\|E_{2,1}^0\right\|_{\mathrm{s}} \le \gamma\ell, \left\|(E_{1,2}^0; E_{2,2}^0)\right\|_{\mathrm{s}} \le \ell$*,* $E_{1,2}^0 \ge 0$ *and* $E_{2,2}^0 \ge 0$ *entry-wise, and* $\|N_{-S}^0\|_\infty \le U$ *and* $\|N_S^0\|_\infty \le 2U \le 1/(16\|(\beta^*)^\dagger\|_\infty)$*. Furthermore, the parameters satisfy that for any* $i \in S$*,*

$$\eta\left(1 + 2r_iR_i\frac{1}{(1-\ell)^2}\frac{B\ell^2}{1-B\ell-\ell} + r_i\frac{2C_1}{k}\left(\alpha+2\rho+\frac{C_1}{k}+\frac{2C_1}{k}\frac{\beta\ell(1-\beta\ell)}{(1-\ell)^2(1-\beta\ell-\ell)}\right)\right) \le \ell \tag{2.8.36}$$

$$r_iR_i\left(2-2\frac{1}{(1-\ell)^2}\frac{\beta\ell^2}{1-\beta\ell-\ell}\right) - r_i\left(\frac{2C_1}{k}\left(\alpha+2\rho+\frac{C_1}{k}\frac{1}{1-\ell}+\frac{2C_1}{k}\frac{\beta\ell(1-\beta\ell)}{(1-\ell)^2(1-\beta\ell-\ell)}\right)\right) \ge 1-\ell \tag{2.8.37}$$

$$h_1 \le \ell, \quad \left(\frac{rR}{(1-\ell)^2}+1\right)(\epsilon+h_1) + (\epsilon+h_2) \le (\gamma-1)\ell \tag{2.8.38}$$

$$\epsilon + h_2 \le (1-\epsilon)(\gamma-1)\ell \tag{2.8.39}$$

$$h_1 + \ell \le (\beta-1)\ell, \quad h_2 + \left(\frac{rR}{(1-\ell)^2}+1\right)\ell \le (\beta-1)\ell \tag{2.8.40}$$

$$3\|(\beta^*)^\dagger\|_\infty(3U+C_\nu) \le \rho < \alpha. \tag{2.8.41}$$

*If we have adversarial noise (Assumption (**N1**)), assume*

$$\epsilon' + U_a \le (1-\epsilon)U, \text{ and } 3\|(\beta^*)^\dagger\|_\infty (2U + U_a + C_\nu) \le \rho < \alpha < 1. \tag{2.8.42}$$

*If we have unbiased noise (Assumption (**N2**)), assume*

$$\epsilon' + U_n \le (1-\epsilon)U. \tag{2.8.43}$$

Finally, let $D = poly(n, m, 1/\delta, 1/\epsilon)$ sufficiently large.

Then with probability at least $1-\delta$, after $\frac{2\ln(\epsilon/(\gamma\ell))}{\ln(1-\eta)} + \frac{\ln(\epsilon'/U)}{\ln(1-\eta)}$ iterations, the output of Algorithm 7 is $\hat{\beta} = \beta^*(\hat{\Sigma}+\hat{E})+\hat{N}$ satisfying

$$(1-\ell)\mathrm{Id} \le \hat{\Sigma} \le u\mathrm{Id}, \quad \|\hat{E}_{1,1}\|_s \le (\gamma-1)\ell, \quad \|\hat{E}_{2,1}\|_s \le (1-\epsilon)(\gamma-1)\ell, \quad \|(\hat{E}_{1,2}; \hat{E}_{2,2})\|_s \le \ell,$$

and $\hat{E}_{1,2} \ge 0$ and $\hat{E}_{2,2} \ge 0$ entry-wise. Furthermore, $\|\hat{N}_{-S}\|_\infty \le U$ and $\|\hat{N}_S\|_\infty \le (1-\epsilon)U$.

Proof of Lemma 72. It follows from Lemma 75 and the conditions (2.8.38) and (2.8.39). □

To prove Lemma 75, we will first consider how E changes after one update step, and then derive the recurrence for all steps in Lemma 75.

2.8.7.1.1 One update step of E

Similarly as before, we focus on one update step first, bounding the change of E. So through out this subsection we will focus on a particular iteration t and omit the superscript (t), while in the next subsection we will put back the superscript.

For analysis, denote $\beta^{(t)}$ as

$$\beta = \beta^*(\Sigma + E) + N$$

where Σ is a diagonal matrix, E is an off-diagonal matrix, and N is the component of β that lies outside the span of β^* (e.g., the noise caused by the noise in the sample).

Recall the following notations:

$$Z = (\Sigma + E)^{-1},$$
$$V = Z - \Sigma^{-1} = \Sigma^{-1} \sum_{i=1}^{\infty} (-E\Sigma^{-1})^i,$$
$$\xi = -\beta^\dagger N Z \gamma^* + \beta^\dagger \nu.$$

Consider the update term $\hat{\mathbb{E}}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right]$ and denote it as

$$\Delta = \hat{\mathbb{E}}\left[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top\right] = \beta^*(\widetilde{\Sigma} + \widetilde{E}) + \widetilde{N}$$

where $\widetilde{\Sigma}$ is a diagonal matrix, $\widetilde{E}$ is an off-diagonal matrix, and N is the component of Δ that lies outside the span of β^*.

Since we now use empirical average, we will have sampling noise. Denote it as

$$N_s = \hat{\mathbb{E}}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top] - \mathbb{E}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top].$$

Then by definition, for $\tilde{f} = \beta^*\gamma^* + \nu$ and $\tilde{f}' = \beta^*(\gamma')^* + \nu'$, we have

$$\begin{aligned}\hat{\mathbb{E}}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top] &= \mathbb{E}[(\tilde{f} - \tilde{f}')(\gamma - \gamma')^\top] + N_s \\ &= \beta^* \underbrace{\mathbb{E}\left[(\gamma^* - (\gamma')^*)(\gamma - \gamma')^\top\right]}_{\widetilde{\Sigma}+\widetilde{E}} + \underbrace{\mathbb{E}\left[(\nu - \nu')(\gamma - \gamma')^\top\right] + N_s}_{\widetilde{N}}.\end{aligned}$$

Recall the definition of $E_{1,1}$, i.e., it is the submatrix of E indexed by $S \times S$. Define $\widetilde{E}_{1,1}$ similarly, i.e., it is the submatrix of $\widetilde{E}$ indexed by $S \times S$. Define $\widetilde{E}_{1,2}, \widetilde{E}_{2,1}$ and $\widetilde{E}_{2,2}$ accordingly. So in the special case when $S = [s]$ where $s = |S|$,

$$\widetilde{E} = \begin{bmatrix} \widetilde{E}_{1,1} & \widetilde{E}_{1,2} \\ \widetilde{E}_{2,1} & \widetilde{E}_{2,2} \end{bmatrix}.$$

We also use the notation M^+ or M^- to denote the positive or negative part of a matrix M.

Lemma 73 (Update $\widetilde{E}_{1,1}$)**.** *Let $\widetilde{E}_{1,1}$ be defined as above. If $\|\xi\|_\infty \le \rho < \alpha < 1$ and $\Sigma \ge (1 - \ell)\mathrm{Id}$, then*

(1). Negative entries:

$$\|\widetilde{E}_{1,1}^-\|_s \le \frac{4C_1^2\|Z\|_s(\|Z\|_s + 1)}{k^2(\alpha - \rho)} + \frac{8C_1(C_1 + 1)\rho\|Z\|_s}{k(\alpha - \rho)}.$$

(2) Positive entries:

$$\|\widetilde{E}^{+}_{1,1}\|_s \le \frac{12C_1(C_1+1)\|Z\|_s}{k^2(\alpha-\rho)}\left(\|Z\|_s + k\rho\right) + 2\max_{j\in[k]}\{\mathbb{E}[(x^*_j)^2]\}\left(\frac{1}{(1-\ell)^2}\|E^{-}_{1,1}\|_s + \frac{\|E\|_s^2}{(1-\ell)^2(1-\ell-\|E\|_s)}\right).$$

Proof of Lemma 73. (1) By Lemma 66, we have

$$\|\widetilde{E}^{-}_{1,1}\|_s \le \max\left\{\frac{4C_1^2\|Z\|_s}{k^2(\alpha-\rho)}\|Z\|_s + \frac{4C_1^2\|Z\|_s}{k^2(\alpha-\rho)}k\rho, \frac{8C_1\rho}{k(\alpha-\rho)}(C_1+1)\|Z\|_s + \frac{2C_1^2}{k^2}\|Z\|_s\right\}.$$

Observe that for $\alpha < 1$,

$$\frac{4C_1^2\|Z\|_s(\|Z\|_s+1)}{k^2(\alpha-\rho)} \ge \max\left\{\frac{4C_1^2\|Z\|_s^2}{k^2(\alpha-\rho)}, \frac{2C_1^2}{k^2}\|Z\|_s\right\}.$$

Moreover,

$$\frac{8C_1\rho}{k(\alpha-\rho)}(C_1+1)\|Z\|_s \ge \frac{4C_1^2\|Z\|_s}{k^2(\alpha-\rho)}k\rho.$$

Therefore,

$$\|\widetilde{E}^{-}_{1,1}\|_s \le \frac{4C_1^2\|Z\|_s}{k^2(\alpha-\rho)} + \frac{8C_1(C_1+1)\rho\|Z\|_s}{k(\alpha-\rho)}.$$

(2) By Lemma 66, when $Z_{i,j} < 0$,

$$\widetilde{E}_{j,i} \le \frac{4C_1^2\|Z^i\|_1}{k^2(\alpha-\rho)}\left(\left|Z_{i,j}\right| + \rho\right).$$

When $Z_{i,j} \ge 0$,

$$\widetilde{E}_{j,i} \le \frac{8C_1\rho}{k(\alpha-\rho)}\left(\frac{C_1\|Z^i\|_1}{k} + Z_{i,j}\right) + 2\mathbb{E}[(\gamma^*_j)^2]Z_{i,j}$$

Consider a fixed i. Let $G = \{j \in S, Z_{i,j} \geq 0\}$ and let $G^c = S - G$. We know that

$$\begin{aligned}
\|[\widetilde{E}^+_{1,1}]_i\|_1 &= \sum_{j\in[k]} [\widetilde{E}^+_{1,1}]_{j,i} \\
&\leq \sum_{j\in G^c} \frac{4C_1^2\|Z^i\|_1}{k^2(\alpha-\rho)} \left(\left|Z_{i,j}\right| + \rho\right) \\
&\quad + \sum_{j\in G} \left(\frac{8C_1\rho}{k(\alpha-\rho)}\left(\frac{C_1\|Z^i\|_1}{k} + Z_{i,j}\right) + 2\mathbb{E}[(\gamma_j^*)^2]Z_{i,j}\right) \\
&\leq \frac{4C_1^2\|Z\|_s}{k^2(\alpha-\rho)} \left(\|Z\|_s + k\rho\right) + \frac{8C_1(C_1+1)\rho}{k(\alpha-\rho)}\|Z\|_s + \sum_{j\in G} 2\mathbb{E}[(\gamma_j^*)^2]Z_{i,j} \\
&\leq \frac{4C_1^2\|Z\|_s^2}{k^2(\alpha-\rho)} + \frac{4C_1^2\|Z\|_s}{k^2(\alpha-\rho)}k\rho + \frac{8C_1(C_1+1)\rho}{k(\alpha-\rho)}\|Z\|_s + \sum_{j\in S} 2\mathbb{E}[(\gamma_j^*)^2]Z_{i,j} \\
&\leq \frac{12C_1(C_1+1)\|Z\|_s}{k^2(\alpha-\rho)} \left(\|Z\|_s + k\rho\right) + \sum_{j\in G} 2\mathbb{E}[(\gamma_j^*)^2]Z_{i,j}.
\end{aligned}$$

A similar bound holds for $\|[\widetilde{E}^+_{1,1}]^i\|_1$.

By the definition of Z, we know that

$$\begin{aligned}
Z &= (\Sigma + E)^{-1} \\
&= \Sigma^{-1} \sum_{m=0}^{\infty} (-E\Sigma^{-1})^m \\
&= \Sigma^{-1} - \Sigma^{-1}E\Sigma^{-1} + \Sigma^{-1} \sum_{m=2}^{\infty} (-E\Sigma^{-1})^m
\end{aligned}$$

Therefore, we know that for $i \neq j$,

$$Z_{i,j} \leq -[\Sigma^{-1}E\Sigma^{-1}]_{i,j} + |\sum_{m=2}^{\infty} \Sigma^{-1}[(-E\Sigma^{-1})^m]_{i,j}|.$$

This implies that

$$\begin{aligned}
\sum_{j\in G} Z_{i,j} &\leq \sum_{j\in G} \left(-[\Sigma^{-1}E\Sigma^{-1}]_{i,j} + \sum_{m=2}^{\infty} \left|\Sigma^{-1}[(-E\Sigma^{-1})^m]_{i,j}\right|\right) \\
&\leq \frac{1}{(1-\ell)^2}\|E^-_{1,1}\|_s + \frac{1}{1-\ell} \frac{\frac{\|E\|_s^2}{(1-\ell)^2}}{1 - \frac{\|E\|_s}{1-\ell}} \\
&\leq \frac{1}{(1-\ell)^2}\|E^-_{1,1}\|_s + \frac{\|E\|_s^2}{(1-\ell)^2(1-\ell-\|E\|_s)}.
\end{aligned}$$

Putting together, we complete the proof. □

Lemma 74 (Update $\widetilde{E}_{2,1}$)**.** *Let $\widetilde{E}_{2,1}$ be defined as above, and suppose $\|\xi\|_\infty \leq \rho < \alpha < 1$, $\Sigma \geq (1-\ell)\mathrm{Id}$ and $E_{1,2} \geq 0$,*

then we have

$$\|\widetilde{E}_{2,1}\|_s \le \frac{12C_1(C_1+1)\|Z\|_s}{k^2(\alpha-\rho)}\left(\|Z\|_s + k\rho\right) + 2\max_{j\in[k]}\{\mathbb{E}[(x_j^*)^2]\}\left(\frac{\|E\|_s^2}{(1-\ell)^2(1-\ell-\|E\|_s)}\right).$$

Proof of Lemma 74. The proof is almost the same as that of Lemma 73, combined with the fact that $E_{1,2} \ge 0$ entry-wise. □

2.8.7.1.2 Recurrence

Recall that

$$\beta = \beta^*(\Sigma + E) + N$$

and recall that $E_{1,1}$ is the submatrix indexed by $S \times S$, and $E_{1,2}, E_{2,1}, E_{2,2}$ are defined according. Recall that M_S denote the submatrix of M formed by columns indexed by S, and let M_{-S} denote the submatrix formed by the other columns.

Lemma 75 (Recurrence)**.** *Suppose the conditions in Lemma 72 hold. Then with probability at least* $1-\delta$*, after* $\frac{2\ln(\epsilon/(\gamma\ell))}{\ln(1-\eta)}$ *iterations,*

$$\begin{aligned}
(1-\ell)\mathrm{Id} &\preceq \Sigma^t,\\
\|(E_{1,1}^t)^-\|_s &\le \epsilon + h_1,\\
\|(E_{1,1}^t)^+\|_s &\le \frac{rR}{(1-\ell)^2}(\epsilon + h_1) + h_2 + \epsilon,\\
\|(E_{2,1}^t)\|_s &\le \epsilon + h_2.
\end{aligned}$$

Also, after $\frac{\ln(\epsilon'/U)}{\ln(1-\eta)}$ *iterations, for both adversarial and unbiased noise,*

$$\left\|N_{-S}^t\right\|_\infty \le U, \quad \left\|N_S^t\right\|_\infty \le (1-\epsilon)U.$$

Proof of Lemma 75. We first prove the following claims by induction.

(1) $(1-\ell)\mathrm{Id} \preceq \Sigma^{(t)}$,

(2)

$$\begin{aligned}
\|(E_{1,1}^{-})^t\|_s &\le \gamma\ell \\
\|(E_{1,1}^{+})^t\|_s &\le \frac{rR}{(1-\ell)^2}\gamma\ell + h_2 \\
\|E_{2,1}^t\|_s &\le \gamma\ell \\
\|E_{1,2}^t\|_s &\le \ell \\
\|E_{2,2}^t\|_s &\le \ell,
\end{aligned}$$

(3) $\|E^t\|_s \le \beta\ell$,

(4) for adversarial noise, $\left\|N_S^t\right\|_\infty \le U + U_a$, and $\|\xi^t\|_\infty \le \rho$; or for unbiased noise, $\left\|N_S^t\right\|_\infty \le U + U_u$.

The basis case for $t = 0$ is trivial by assumptions. Now assume they are true for iteration t and show that they are true for iteration $t + 1$.

(1) By the update of Σ, we have

$$\Sigma^{t+1} = (1-\eta)\Sigma^t + \eta r\widetilde{\Sigma}^t.$$

To lower bound $\Sigma_{i,i}^{t+1}$, we will consider two cases, $\Sigma_{i,i}^t \ge 1$ and $\Sigma_{i,i}^t \le 1$.

For $\Sigma_{i,i}^t \ge 1$, by Lemma 65,

$$\begin{aligned}
\widetilde{\Sigma}_{i,i} &\ge \mathbb{E}\left[(x_i^*)^2\right]\left(2\Sigma_{i,i}^{-1} - 2\left|V_{i,i}\right|\right) - \frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}\Sigma_{i,i}^{-1} + \frac{2C_1}{k}\left\|[V]^i\right\|_1\right) \\
&\ge -2R_i\left|V_{i,i}\right| - \left(\frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}\Sigma_{i,i}^{-1} + \frac{2C_1}{k}\left\|[V]^i\right\|_1\right)\right).
\end{aligned}$$

Hence,

$$\begin{aligned}
\Sigma_{i,i}^{t+1} &\ge (1-\eta)\Sigma_{i,i}^t - \eta\left(2r_iR_i\left|V_{i,i}^t\right| + r_i\left(\frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}(\Sigma_{i,i}^t)^{-1} + \frac{2C_1}{k}\left\|\left[V^t\right]^i\right\|_1\right)\right)\right) \\
&\ge 1 - \eta\left(1 + 2r_iR_i\left|V_{i,i}^t\right| + r_i\frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k} + \frac{2C_1}{k}\left\|\left[V^t\right]^i\right\|_1\right)\right) \\
&\ge 1 - \eta\left(1 + 2r_iR_i\frac{1}{(1-\ell)^2}\frac{\beta\ell^2}{1-\beta\ell-\ell} + r_i\frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k} + \frac{2C_1}{k}\frac{\beta\ell(1-\beta\ell)}{(1-\ell)^2(1-\beta\ell-\ell)}\right)\right).
\end{aligned}$$

where we use the bound on V^t. By condition (2.8.36), the claim follows.

For $\Sigma_{i,i}^t \le 1$, again by Lemma 65,

$$\widetilde{\Sigma}_{i,i}^t \ge \mathbb{E}\left[(x_i^*)^2\right]\left(2 - 2\left|V_{i,i}^t\right|\right) - \left(\frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}(\Sigma_{i,i}^t)^{-1} + \frac{2C_1}{k}\left\|\left[V^t\right]^i\right\|_1\right)\right).$$

Hence,

$$\begin{aligned}
\Sigma_{i,i}^{t+1} &= (1-\eta)\Sigma_{i,i}^t + \eta r \widetilde{\Sigma}_{i,i}^t \\
&\geq (1-\eta)(1-\ell) \\
&\quad + \eta\left(r_i R_i\left(2 - 2\left|V_{i,i}^t\right|\right) - r_i\left(\frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}(\Sigma_{i,i}^t)^{-1} + \frac{2C_1}{k}\left\|\left[V^t\right]^i\right\|_1\right)\right)\right) \\
&\geq (1-\eta)(1-\ell) + \eta r_i R_i\left(2 - 2\frac{1}{(1-\ell)^2}\frac{\beta\ell^2}{1-\beta\ell-\ell}\right) \\
&\quad - \eta r_i\left(\frac{2C_1}{k}\left(\alpha + 2\rho + \frac{C_1}{k}\frac{1}{1-\ell} + \frac{2C_1}{k}\frac{\beta\ell(1-\beta\ell)}{(1-\ell)^2(1-\beta\ell-\ell)}\right)\right).
\end{aligned}$$

By condition (2.8.37), the claim follows.

(2) By Lemma 73,

$$\begin{aligned}
\|(\widetilde{E}_{1,1}^{t+1})^-\|_s &\leq \frac{8C_1(C_1+1)\rho\|Z^t\|_s}{k(\alpha-\rho)} + \frac{4C_1^2\|Z^t\|_s(\|Z^t\|_s+1)}{k^2(\alpha-\rho)}, \\
\|(\widetilde{E}_{1,1}^{t+1})^+\|_s &\leq \frac{R}{(1-\ell)^2}\|(E_{1,1}^-)^t\|_s + \frac{R\|E^t\|_s^2}{(1-\ell)^2(1-\ell-\|E^t\|_s)} \\
&\quad + \frac{12C_1(C_1+1)\|Z^t\|_s}{k^2(\alpha-\rho)}\left(\left\|Z^t\right\|_s + k\rho\right).
\end{aligned}$$

By the update rule, we have

$$\begin{aligned}
\|(E_{1,1}^{t+1})^-\|_s &\leq (1-\eta)\|(E_{1,1}^t)^-\|_s \\
&\quad + r\eta\frac{8C_1(C_1+1)\rho}{(1-\ell-\beta\ell)k(\alpha-\rho)} + \frac{4C_1^2}{(1-\ell-\beta\ell)k^2(\alpha-\rho)}\left(\frac{1}{(1-\ell-\beta\ell)}+1\right)r\eta, \\
&\leq (1-\eta)\|(E_{1,1}^t)^-\|_s + \eta h_1 \qquad (2.8.44) \\
\|(E_{1,1}^{t+1})^+\|_s &\leq (1-\eta)\|(E_{1,1}^t)^+\|_s + r\eta\frac{R}{(1-\ell)^2}\|(E_{1,1}^t)^-\|_s \\
&\quad + r\eta\frac{R\beta^2\ell^2}{(1-\ell)^2(1-\ell-\beta\ell)} \\
&\quad + \frac{12C_1(C_1+1)}{k^2(\alpha-\rho)(1-\ell-\beta\ell)}\left(\frac{1}{1-\ell-\beta\ell}+k\rho\right)r\eta \\
&\leq (1-\eta)\|(E_{1,1}^t)^+\|_s + r\eta\frac{R}{(1-\ell)^2}\|(E_{1,1}^t)^-\|_s + \eta h_2 \qquad (2.8.45)
\end{aligned}$$

where we use $\left\|E^t\right\|_s \leq \beta\ell$ and $\|Z^t\|_s \leq \frac{1}{1-\ell-\beta\ell}$.

The claim on $\|(E_{1,1}^{t+1})^-\|_s$ follows from (2.8.44) and the condition (2.8.38).

For $\|(E_{1,1}^{t+1})^+\|_s$, by induction (2.8.45) becomes

$$\|(E_{1,1}^{t+1})^+\|_s \le (1-\eta)\|(E_{1,1}^{t})^+\|_s + r\eta\frac{R}{(1-\ell)^2}\gamma\ell + \eta h_2 \le \frac{rR}{(1-\ell)^2}\gamma\ell + h_2.$$

Now we consider $\|(E_{2,1}^{t+1})\|_s$. By Lemma 74,

$$\begin{aligned}\|(E_{2,1}^{t+1})\|_s &\le (1-\eta)\|(E_{2,1}^{(t)})\|_s \\ &\quad + r\eta\frac{R\beta^2\ell^2}{(1-\ell)^2(1-\ell-\beta\ell)} \\ &\quad + \frac{12C_1(C_1+1)}{k^2(\alpha-\rho)(1-\ell-\beta\ell)}\left(\frac{1}{1-\ell-\beta\ell}+k\rho\right)r\eta \\ &= (1-\eta)\|(E_{2,1}^{t})\|_s + \eta h_2 \\ &\le \gamma\ell\end{aligned} \tag{2.8.46}$$

where the last line follows by condition (2.8.39) and induction.

Finally, clearly we have $\left\|E_{1,2}^{t+1}\right\|_s \le \ell$ and $\left\|E_{2,2}^{t+1}\right\|_s \le \ell$, since they are not updated.

(3) Note that (2.8.44) (2.8.45) hold for all iterations up to $t+1$. Then by Lemma 80, we have

$$\begin{aligned}&\|(E_{1,1}^{t+1})^-\|_s + \|(E_{1,1}^{t+1})^+\|_s \\ &\le \max\left\{\|(E_{1,1}^{0})^-\|_s + \|(E_{1,1}^{0})^+\|_s, \|(E_{1,1}^{0})^+\|_s + h_1, h_2 + \left(\frac{rR}{(1-\ell)^2}+1\right)\|(E_{1,1}^{0})^-\|_s, h_2 + \left(\frac{rR}{(1-\ell)^2}+1\right)h_1\right\}.\end{aligned}$$

Since $h_1 \le \ell$ and $h_2 \le \ell$ by (2.8.38)(2.8.39), and $\|(E_{1,1}^{0})^-\|_s + \|(E_{1,1}^{0})^+\|_s \le \ell$ by assumption, we have

$$\|(E_{1,1}^{t+1})^-\|_s + \|(E_{1,1}^{t+1})^+\|_s \le \max\left\{\ell + h_1, h_2 + \left(\frac{rR}{(1-\ell)^2}+1\right)\ell\right\}. \tag{2.8.47}$$

Then we have by condition (2.8.40),

$$\|(E_{1,1}^{t+1})^-\|_s + \|(E_{1,1}^{t+1})^+\|_s \le (\beta-1)\ell, \quad \|E^{t+1}\|_s \le \beta\ell.$$

(4) Finally, we consider the noise. We first consider the adversarial noise. Set the sample size D to be large enough, so that by Lemma 67, we have

$$\left|\widetilde{N}_{i,j}^{t}\right| \le \frac{4C_\nu C_1}{(1-2\ell)^2 k(\alpha-\rho)} + \left|[\widetilde{N}_{s}^{t}]_{i,j}\right| \le \frac{8C_\nu C_1}{k(\alpha-\rho)}$$

and thus

$$\left\|N^{t+1}\right\|_\infty \le (1-\eta)\left\|N^t\right\|_\infty + \eta\frac{8rC_\nu C_1}{\alpha-\rho}. \tag{2.8.48}$$

Then for any $t \ge 0$,

$$\left\|N^t\right\|_\infty \le \left\|N^0\right\|_\infty + \frac{8rC_\nu C_1}{\alpha-\rho} \le U + \frac{8rC_\nu C_1}{\alpha-\rho} \le 2U + U_a$$

where the last inequality is by the definition of U_a. On the other hand, by Lemma 69, we have

$$\begin{aligned}
\|\xi^{(t)}\|_\infty &\le 3\|(\beta^*)^\dagger\|_\infty(\|N^t\|_\infty + C_\nu) \\
&\le 3\|(\beta^*)^\dagger\|_\infty\left(2U + \frac{8rC_\nu C_1}{\alpha-\rho} + C_\nu\right) \\
&\le \rho
\end{aligned}$$

where the last inequality is due to condition (2.8.42).

We now consider the unbiased noise, where the proof is similar. Set the sample size N to be large enough, so that by Lemma 67, we have

$$\begin{aligned}
\left|\widetilde{N}^t_{i,j}\right| &\le \frac{2C_1C_\nu\rho'(1+\|\beta^\dagger N^{(t)}\|_\infty)}{(1-2\ell)k(\alpha-\rho')} + \left|[N_s]_{i,j}\right| \\
&\le \frac{8C_1C_\nu^2\left\|(\beta^*)^\dagger\right\|_\infty}{(1-2\ell)k(\alpha-2C_\nu\left\|(\beta^*)^\dagger\right\|_\infty)} + \left|[N_s]_{i,j}\right| \\
&\le \frac{10C_1C_\nu^2\left\|(\beta^*)^\dagger\right\|_\infty}{(1-2\ell)k(\alpha-2C_\nu\left\|(\beta^*)^\dagger\right\|_\infty)},
\end{aligned}$$

and thus

$$\left\|N_S^{t+1}\right\|_\infty \le (1-\eta)\left\|N_S^t\right\|_\infty + \eta\frac{10rC_1C_\nu^2\left\|(\beta^*)^\dagger\right\|_\infty}{(1-2\ell)(\alpha-2C_\nu\left\|(\beta^*)^\dagger\right\|_\infty)}. \tag{2.8.49}$$

Then for any $t \ge 0$,

$$\left\|N_S^t\right\|_\infty \le \left\|N_S{}^0\right\|_\infty + \frac{10rC_1C_\nu^2\left\|(\beta^*)^\dagger\right\|_\infty}{(1-2\ell)(\alpha-2C_\nu\left\|(\beta^*)^\dagger\right\|_\infty)} \le 2U + \frac{10rC_1C_\nu^2\left\|(\beta^*)^\dagger\right\|_\infty}{(1-2\ell)(\alpha-2C_\nu\left\|(\beta^*)^\dagger\right\|_\infty)} \le 2U + U_n$$

where the last inequality is by the definition of U_n. This completes the proof for the claims.

Now, after proving the claims, we are ready to prove the last statement of the lemma. First, by (2.8.44) and

Lemma 81, we have that after $\frac{\ln(\epsilon/(\gamma\ell))}{\ln(1-\eta)}$ iterations,

$$\|(E_{1,1}^t)^-\|_s \le \epsilon + h_1.$$

Now (2.8.45) becomes

$$\|(E_{1,1}^{t+1})^+\|_s \le (1-\eta)\|(E_{1,1}^t)^+\|_s + r\eta\frac{R}{(1-\ell)^2}(\epsilon + h_1) + \eta h_2 \tag{2.8.50}$$

After an additional $\frac{\ln(\epsilon/(\gamma\ell))}{\ln(1-\eta)}$ iterations, by Lemma 81,

$$\|(E_{1,1}^t)^+\|_s \le \frac{rR}{(1-\ell)^2}(\epsilon + h_1) + h_2 + \epsilon$$

Similarly, Lemma 81 and (2.8.46), after $\frac{\ln(\epsilon/(\gamma\ell))}{\ln(1-\eta)}$ iterations,

$$\|(E_{2,1}^t)\|_s \le \epsilon + h_2.$$

$\left\|N_{-S}^t\right\|_\infty$ does not change since it is not updated. Now consider $\left\|N_S^t\right\|_\infty$.

For the adversarial noise, by (2.8.48) and Lemma 81, after $\frac{\ln(\epsilon'/U)}{\ln(1-\eta)}$ iterations,

$$\left\|N_S^t\right\|_\infty \le \epsilon' + \frac{8rC_\nu C_1}{\alpha-\rho} \le (1-\epsilon)U$$

where the last inequality is due to condition (2.8.42).

For the unbiased noise, by (2.8.49) and Lemma 81, after $\frac{\ln(\epsilon'/U)}{\ln(1-\eta)}$ iterations,

$$\left\|N_S^t\right\|_\infty \le \epsilon' + \frac{10rC_1C_\nu^2\left\|(\beta^*)^\dagger\right\|_\infty}{(1-2\ell)(\alpha - 2C_\nu\left\|(\beta^*)^\dagger\right\|_\infty)} \le (1-\epsilon)U$$

where the last inequality is due to condition (2.8.43).

This completes the proof. □

2.8.7.2 Equilibration: Rescale

The input of of Algorithm 8 can be written as $\beta^0 = \beta^*(\Sigma^0 + E^0) + N^0$. The output $\hat{\beta}$ can be written as $\hat{\beta} = (\beta^* D)(\hat{\Sigma} + \hat{E}) + \hat{N}$ where $\hat{\Sigma}$ is diagonal, and $\hat{E}$ is off diagonal, and D is a diagonal matrix with $D_{i,i} = \frac{1}{1-\epsilon}$ for $i \in S$ and the rest being 1. Recall that for a matrix M, let $M_{1,1}$ denote the submatrix of M indexed by $S \times S$, and define $M_{1,2}, M_{2,1}$ and $M_{2,2}$ accordingly. Also recall that M_S denote the submatrix of M formed by columns indexed by S, and let M_{-S} denote the

submatrix formed by the other columns.

Lemma 76 (Main: Rescale)**.** *Let $\beta^0 = \beta^*(\Sigma^0 + E^0) + N^0$ satisfies the condition in Lemma 72 and ϵ be defined as in Lemma 72. Then the output of Algorithm 8 is $\hat{\beta} = (\beta^* D)(\hat{\Sigma} + \hat{E}) + \hat{N}$ satisfying*

$$(1-\ell)\mathrm{Id} \le \hat{\Sigma}, \quad \|\hat{E}_{1,1}\|_s \le (\gamma-1)\ell, \quad \|\hat{E}_{2,1}\|_s \le (\gamma-1)\ell, \quad \|(\hat{E}_{1,2}, \hat{E}_{2,2})\|_s \le \ell, \quad \|\hat{N}_S\|_\infty \le U, \quad \|\hat{N}_{-S}\|_\infty \le U.$$

Moreover, $\hat{E}_{1,2} \ge 0$ and $\hat{E}_{2,2} \ge 0$ entry-wise.

Proof of Lemma 76. Note that $\tilde{\beta} = \beta^*(\widetilde{\Sigma} + \widetilde{E}) + \widetilde{N}$ for a diagonal matrix $\widetilde{\Sigma}$, off-diagonal matrix $\widetilde{E}$ and error matrix $\widetilde{N}$. By lemma 72, we have $\widetilde{\Sigma} \ge (1-\ell)\mathrm{Id}$, error matrix $\|\widetilde{N}_S\|_\infty \le (1-\epsilon)U$ and

$$\|\widetilde{E}_{1,1}\|_s \le (\gamma-1)\ell, \quad \|\widetilde{E}_{2,1}\|_s \le (1-\epsilon)(\gamma-1)\ell, \quad \|(\widetilde{E}_{1,2}; \widetilde{E}_{2,2})\|_s \le \ell$$

and $\widetilde{E}_{1,2} \ge 0$ and $\widetilde{E}_{2,2} \ge 0$ entry-wise.

Therefore, by the rescaling rule:

$$\begin{aligned}\hat{\beta} = \tilde{\beta} D &= \beta^*(\widetilde{\Sigma} + \widetilde{E})D + \widetilde{N}D \\ &= \beta^* D(\widetilde{\Sigma} + D^{-1}\widetilde{E}D) + \widetilde{N}D.\end{aligned}$$

Therefore, $\hat{\Sigma} = \widetilde{\Sigma} \ge (1-\ell)\mathrm{Id}$, $\|\hat{N}_S\|_\infty \le \frac{1}{1-\epsilon}\|\widetilde{N}_S\|_\infty \le U$. $\|\hat{N}_{-S}\|_\infty = \|\widetilde{N}_{-S}\|_\infty \le U$ since it is not updated.

For the $\hat{E}$ term, denote $D_1 = \mathsf{Diag}\left(\frac{1}{1-\epsilon}, \ldots, \frac{1}{1-\epsilon}\right) \in \mathbb{R}^{s\times s}$. We know that

$$\begin{aligned}\hat{E}_{1,1} &= D_1^{-1}\widetilde{E}_{1,1}D_1 = \widetilde{E}_{1,1} \\ \hat{E}_{2,1} &= \widetilde{E}_{2,1}D_1 = \frac{1}{1-\epsilon}\widetilde{E}_{2,1} \\ \hat{E}_{1,2} &= D_1^{-1}\widetilde{E}_{1,2} = (1-\epsilon)\widetilde{E}_{1,2} \\ \hat{E}_{2,2} &= \widetilde{E}_{2,2}.\end{aligned}$$

This leads to

$$\|\hat{E}_{1,1}\|_s \le (\gamma-1)\ell, \quad \|\hat{E}_{2,1}\|_s \le (\gamma-1)\ell, \quad \|(\hat{E}_{1,2}, \hat{E}_{2,2})\|_s \le \ell,$$

with $\hat{E}_{1,2}, \hat{E}_{2,2} \ge 0$. This completes the proof. □

2.8.7.3 Equilibration: Main algorithm

Lemma 77 (Main: Equilibration)**.** *Suppose the conditions in Lemma 76 each time Algorithm 8. Additionally, there exists constant* $0 < b < 1$, $\kappa > 1$ *and* $u > 1$ *such that* $b\kappa > 1$ *such that the initial* $\lambda \geq \max_{i\in[k]} \mathbb{E}[(x_i^*)^2]/b$, *and the initial* $\Sigma \preceq u\mathrm{Id}$. *Furthermore, for any* $\lambda \geq \min_{i\in[k]} \mathbb{E}[(x_i^*)^2]/\kappa$,

$$\left(\frac{1}{1-\ell} + h_6\right)^2 b\lambda + h_5^2 b\kappa\lambda + h_3 \leq \left(1 - \frac{1}{100}\right)\lambda, \tag{2.8.51}$$

$$\left(\frac{1}{u} - h_6\right)^2 (1-\epsilon) b\kappa\lambda - h_5^2 b\kappa\lambda - h_4 \geq \left(1 + \frac{1}{100}\right)\lambda, \quad \frac{1}{u} > h_6 \tag{2.8.52}$$

$$h_3 \leq \frac{1}{200} \min_{i\in[k]} \mathbb{E}[(x_i^*)^2], \tag{2.8.53}$$

$$h_4 \leq \frac{1}{200} \min_{i\in[k]} \mathbb{E}[(x_i^*)^2], \tag{2.8.54}$$

where

$$\begin{aligned}
h_3 &= \frac{C_1^2}{k^2} h_5 \left(h_5 + \frac{2}{1-\ell}\right), \\
h_4 &= \frac{C_1^2}{k^2} h_5 \left(h_5 + \frac{2}{1-\ell}\right) + \frac{2(\alpha+\rho)C_1}{k(1-\ell)}, \\
h_5 &= \frac{(\gamma+1)\ell(1-(\gamma+1)\ell)}{(1-\ell)^2(1-(\gamma+2)\ell)}, \\
h_6 &= \frac{(\gamma+1)\ell^2}{(1-\ell)^2(1-(\gamma+2)\ell)}.
\end{aligned}$$

Finally, set $N = poly(1/\min_{i\in[k]} \mathbb{E}[(\gamma_i^*)^2], k, 1/\delta)$ *large enough.*

Then with probability at least $1-\delta$, *the following hold. During the execution of the algorithm, for any* $j \in S$,

$$\left(\left(\frac{1}{u} - h_6\right)^2 - \kappa h_5^2 - \frac{1}{100}\right) \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j})^2} \leq M_j \leq \left(\left(\frac{1}{1-\ell} + h_6\right)^2 + \kappa h_5^2 + \frac{1}{100}\right) \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j})^2}.$$

Furthermore, the output of Algorithm 9 is $\beta = \beta^* D(\Sigma + E) + N$ *where* Σ *is diagonal and* $(1-\ell)\mathrm{Id} \preceq \Sigma$, E *is off diagonal and* $\|E\|_s \leq \gamma\ell$, N *satisfies* $\|N\|_\infty \leq 2U$, *and*

$$\frac{\max_{i\in[k]} \frac{1}{D_{i,i}^2} \mathbb{E}[(\gamma_i^*)^2]}{\min_{j\in[k]} \frac{1}{D_{j,j}^2} \mathbb{E}[(\gamma_j^*)^2]} \leq \kappa.$$

Proof of Lemma 77. We prove the lemma by induction. For notational convenience, let us introduce a counter (p) denoting the number of times the inner while cycle has been executed, and denote β as $\beta^{(p)}$. Recall that for a matrix $M \in \mathbb{R}^{k\times k}$ and index set $S \subseteq [k]$, let $M_{1,1}$ denote the submatrix indexed by $S \times S$, and $M_{1,2}, M_{2,1}$ and $M_{2,2}$ are defined accordingly. Also, let M_S denote the submatrix formed by the columns indexed by S, and M_{-S} the submatrix formed

by the other columns.

Our inductive claims are as follows. At the beginning of each inner while cycle,

$$\beta^p = \beta^* D^p \left(\Sigma^p + E^p\right) + N^p$$

where D^p and Σ^p are diagonal, E^p are off diagonal satisfying

(1) $(1-\ell)\mathrm{Id} \preceq \Sigma^p$,

(2) $E^p_{1,2} \geq 0$ and $E^p_{2,2} \geq 0$ entry-wise and

$$\left\|E^p_{1,1}\right\|_{\mathrm{s}} \leq \gamma\ell,$$
$$\left\|E^p_{2,1}\right\|_{\mathrm{s}} \leq \gamma\ell,$$
$$\left\|(E^p_{1,2}; E^p_{2,2})\right\|_{\mathrm{s}} \leq \ell,$$

(3) $N^p_{-S} \leq U$ and $N^p_S \leq 2U$,

(4) We have

(a) When $\mathbb{E}[(\gamma^*_j)^2] < b\lambda^p$, $j \notin S$, then $M_j \leq \lambda^p$,

(b) When $\mathbb{E}[(\gamma^*_j)^2] \geq (1-\epsilon)b\kappa\lambda^p$, $j \notin S$, then $M_j > \lambda^p$,

and consequently,

(c) $\forall i \in S, b\lambda^p \leq \frac{\mathbb{E}[(\gamma^*_i)^2]}{\left(D^p_{i,i}\right)^2}$,

(d) $\forall i \in [k], \frac{\mathbb{E}[(\gamma^*_i)^2]}{\left(D^p_{i,i}\right)^2} \leq b\kappa\lambda^p$.

The claims are trivially true at initialization, so we proceed to the induction. Assume the claim is true at time p, we proceed to show it is true at time $p+1$.

First, consider (1), (2) and (3). By Lemma 76, after applying the rescaling algorithm, $(1-\ell)\mathrm{Id} \preceq \Sigma^p$ and

$$\|E^p_{1,1}\|_s \leq (\gamma-1)\ell, \quad \|E^p_{2,1}\|_s \leq (\gamma-1)\ell, \quad \|(E^p_{1,2}, E^p_{2,2})\|_s \leq \ell, \quad \|N^p_S\|_\infty \leq U, \quad \|N^p_{-S}\|_\infty \leq U.$$

Moreover, $E^p_{1,2} \geq 0$ and $E^p_{2,2} \geq 0$ entry-wise. Observe that when moving from time p to $p+1$, potentially the algorithm includes new elements in S. Then

$$\|E^{p+1}_{1,1}\|_s \leq \|E^p_{1,1}\|_s + \max\{\|E^p_{2,1}\|_s, \|E^p_{1,2}\|_s\} \leq (\gamma-1)\ell + \ell = \gamma\ell$$

Where the last inequality used the fact that $\gamma < 2$. Similarly,

$$\|E_{2,1}^{p+1}\|_s \le \|E_{2,1}^{p}\|_s + \|E_{2,2}^{p}\|_s \le (\gamma - 1)\ell + \ell = \gamma\ell.$$

Also, $\|(E_{1,2}^{p+1}, E_{2,2}^{p+1})\|_s \le \|(E_{1,2}^{p}, E_{2,2}^{p})\|_s \le \ell$, and $(E_{1,2}^{p+1}, E_{2,2}^{p+1}) \ge 0$ entry-wise. Furthermore, $\|N_{-S}^{p+1}\|_\infty \le \|N_{-S}^{p}\|_\infty \le U$ and

$$\|N_{S}^{p+1}\|_\infty \le \|N_{S}^{p}\|_\infty + \|N_{-S}^{p}\|_\infty \le 2U.$$

Hence, (1), (2) and (3) are also true at time $(p+1)$.

Finally, we proceed to (4). Since (a)(b) are true at time p, (c)(d) are true at time $p+1$. [8] Furthermore, when $\lambda \le \min_{i\in[k]} \mathbb{E}[(\gamma_i^*)^2]/\kappa$, it is guaranteed that all $[k] \subseteq S$, so we only need to prove that when $\lambda \ge \min_{i\in[k]} \mathbb{E}[(\gamma_i^*)^2]/\kappa$, (a)(b) are also true at time $p+1$.

To prove (a)(b) are true at time $p+1$, we will use Lemma 78. Note that since β has been scaled, so $\beta^* D$ should be regarded as the ground truth matrix β^* in Lemma 78. We first make sure its assumption is satisfied. First, $\|N\|_\infty \le 3U$ and $\left\|(\beta^* D)^\dagger\right\|_\infty \le \left\|(\beta^*)^\dagger\right\|_\infty$. By Lemma 69 and condition (2.8.41), the assumption in Lemma 78 is satisfied.

We are now ready to prove (a). By Lemma 78,

$$\mathbb{E}[x_j^2] \le \left(\Sigma_{j,j}^{-1} + |V_{j,j}|\right)^2 \frac{\mathbb{E}[(\gamma_j^*)^2]}{D_{j,j}^{p+1}} + \|[V]^j\|_2^2 \max_{m\in[k]} \frac{\mathbb{E}[(\gamma_m^*)^2]}{D_{m,m}^{p+1}} + \|[V]^j\|_1 \left(\|[V]^j\|_1 + 2\Sigma_{j,j}^{-1}\right) \frac{C_1^2}{k^2}.$$

By Lemma 68, $|V_{j,j}| \le h_6$, $\|[V]^j\|_2^2 \le \|[V]^j\|_1^2 \le h_5^2$, so

$$\mathbb{E}[\gamma_j^2] \le \left(\frac{1}{1-\ell} + h_6\right)^2 \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j}^{p+1})^2} + h_5^2 \max_{m\in[k]} \frac{\mathbb{E}[(\gamma_k^*)^2]}{(D_{k,k}^{p+1})^2} + h_3.$$

By (d), $\max_{m\in[k]} \frac{\mathbb{E}[(\gamma_m^*)^2]}{(D_{m,m}^{p+1})^2} \le b\kappa\lambda$, so for any $j \notin S$ with $\mathbb{E}[(\gamma_j^*)^2] = \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j}^{(p+1)})^2} < b\lambda$, we have

$$\mathbb{E}[\gamma_j^2] \le \left(\frac{1}{1-\ell} + h_6\right)^2 b\lambda + h_5^2 b\kappa\lambda + h_3.$$

By using large enough sample, with high probability, the empirical estimation

$$\hat{\mathbb{E}}[\gamma_j^2] \le \mathbb{E}[\gamma_j^2] + \frac{1}{100}\lambda \le \lambda$$

where the last step is by condition (2.8.51).

[8] Note that in (b), the factor $(1-\epsilon)$ is needed to ensure (d) is true at time $p+1$.

As for (b), by Lemma 78 we have

$$\mathbb{E}[\gamma_j^2] \geq \left(\Sigma_{j,j}^{-1} - |V_{j,j}|\right)^2 \frac{\mathbb{E}[(x_j^*)^2]}{(D_{j,j}^{p+1})^2} - \|[V]^j\|_2^2 \max_{k\in[n]} \frac{\mathbb{E}[(x_k^*)^2]}{(D_{k,k}^{p+1})^2} - \left(\frac{C_1^2}{k^2}\|[V]^j\|_1(\|[V]^j\|_1 + 2\Sigma_{j,j}^{-1}) + \frac{2(\alpha+\rho)C_1}{k}\Sigma_{j,j}^{-1}\right)$$
$$\geq \left(\frac{1}{u} - h_6\right)^2 \frac{\mathbb{E}[(x_j^*)^2]}{(D_{j,j}^{p+1})^2} - h_5^2 \max_{m\in[k]} \frac{\mathbb{E}[(\gamma_m^*)^2]}{(D_{m,m}^{p+1})^2} - h_4.$$

The last step uses that $\Sigma_{j,j}^{-1} \leq u$, which is by the initial condition assumed and that it is not updated for $j \notin S$. Putting in the bound that $\frac{\mathbb{E}[(\gamma_m^*)^2]}{(D_{m,m}^{p+1})^2} \leq b\kappa\lambda$, then for any $j \notin S$ with $\mathbb{E}[(\gamma_j^*)^2] = \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j}^{p+1})^2} \geq (1-\epsilon)b\kappa\lambda$, we have

$$\mathbb{E}[\gamma_j^2] \geq \left(\frac{1}{u} - h_6\right)^2 (1-\epsilon)b\kappa\lambda - h_5^2 b\kappa\lambda - h_4.$$

Again, use large enough sample to ensure that with high probability

$$\tilde{\mathbb{E}}[\gamma_j^2] \geq \mathbb{E}[\gamma_j^2] - \frac{1}{100}\lambda \geq \lambda$$

where the last step follows from condition (2.8.52). This completes the proof of the induction.

We now prove the statements of the lemma. The statement about the output follows from the above claims. What is left is to prove that $M_j (j \in S)$ approximates $\mathbb{E}[(x_j^*)^2]$ well. Since M_j for $j \in S$ is updated along with $D_{j,j}$, we only need to check the right after adding j to S, the statement holds. Suppose the time point is p, we have

$$\mathbb{E}[\gamma_j^2] \leq \left(\frac{1}{1-\ell} + h_6\right)^2 \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j}^{p})^2} + h_5^2 \max_{m\in[k]} \frac{\mathbb{E}[(\gamma_m^*)^2]}{(D_{m,m}^{p})^2} + h_3.$$

Since j is in S, by the claims (c)(d) we have

$$\max_{m\in[k]} \frac{\mathbb{E}[(k_m^*)^2]}{(D_{m,m}^{p})^2} \leq b\kappa\lambda^p \leq \kappa \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j}^{p})^2}.$$

Since D is large enough so that

$$\mathbb{E}[\gamma_j^2] \leq \mathbb{E}[(\gamma_j^*)^2]\left(1 + \frac{1}{200}\right).$$

Combined these with the condition (2.8.53), we have

$$M_j \leq \left(\left(\frac{1}{1-\ell} + h_6\right)^2 + \kappa h_5^2 + \frac{1}{100}\right) \frac{\mathbb{E}[(\gamma_j^*)^2]}{(D_{j,j})^2}.$$

The upper bound on M_j can be bounded similarly. This completes the proof of the lemma. □

The following is the lemma used in the proof of Lemma 77.

Lemma 78 (Estimate of feature weight)**.** *Suppose* $|\xi_i| \le \rho < \alpha$ *for any example and every* $i \in [k]$*, and suppose* $\Sigma \ge \frac{1}{2}\mathrm{Id}$*. Then*

$$\begin{aligned}\mathbb{E}[\gamma_i^2] &\ge \left(\Sigma_{i,i}^{-1} - |V_{i,i}|\right)^2 \mathbb{E}[(\gamma_i^*)^2] - \|[V]^i\|_2^2 \max_{j\in[k]} \mathbb{E}[(x_j^*)^2] \\ &\quad - \left(\frac{C_1^2}{k^2}\|[V]^i\|_1(\|[V]^i\|_1 + 2\Sigma_{i,i}^{-1}) + \frac{2(\alpha+\rho)C_1}{k}\Sigma_{i,i}^{-1}\right) \\ \mathbb{E}[\gamma_i^2] &\le \left(\Sigma_{i,i}^{-1} + |V_{i,i}|\right)^2 \mathbb{E}[(\gamma_i^*)^2] + \|[V]^i\|_2^2 \max_{j\in[k]} \mathbb{E}[(\gamma_j^*)^2] + \|[V]^i\|_1\left(\|[V]^i\|_1 + 2\Sigma_{i,i}^{-1}\right)\frac{C_1^2}{k^2}.\end{aligned}$$

Proof of Lemma 78. By the decoding rule,

$$\begin{aligned}\gamma_i &= \left[\phi_\alpha(\beta^\dagger[\beta^*\gamma^* + \nu])\right]_i \\ &= \left[\phi_\alpha\left(\left(\Sigma^{-1} + V\right)\gamma^* + \xi\right)\right]_i.\end{aligned}$$

Let $[V]^i = v$ and $\Sigma_{i,i}^{-1} = \sigma$, then we can rewrite above as

$$\gamma_i = \phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)$$

which implies that

$$\sigma\gamma_i^* + \langle v, \gamma^*\rangle - \rho - \alpha \le x_i \le \left|\sigma\gamma_i^* + \langle v, \gamma^*\rangle\right|. \tag{2.8.55}$$

First, consider the lower bound.

$$\mathbb{E}[\gamma_i^2] \ge \mathbb{E}\left[\left(\sigma\gamma_i^* + \langle v, \gamma^*\rangle - \rho - \alpha\right)\phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)\right]$$

The following simple lemma is useful.

Claim 1. *Let* χ *be a variable such that* $|\chi| \le \alpha$*, then for every* $w \in \mathbb{R}^k$*,* $i \in [k]$*,*

$$\mathbb{E}[\gamma_i^*\phi_\alpha(\langle w, x^*\rangle + \chi)] \le |w_i|\mathbb{E}[(\gamma_i^*)^2] + \frac{C_1^2}{k^2}\sum_{j\ne i}|w_j| \tag{2.8.56}$$

$$\le |w_i|\mathbb{E}[(\gamma_i^*)^2] + \frac{C_1^2}{k^2}\|w\|_1. \tag{2.8.57}$$

Proof. The proof is a direct observation that when $|\chi| < \alpha$,

$$\phi_\alpha(\langle w, x\rangle + \chi) \le |\langle w, x\rangle| \le \langle |w|, x\rangle$$

where $|w|$ is the entry wise absolute value. □

Therefore, we can obtain the following bounds.

(1). By (2.8.13) in Lemma 65, we have

$$\mathbb{E}[\gamma_i^* \phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)] \ge \Sigma_{i,i}^{-1}\mathbb{E}\left[(\gamma_i^*)^2\right] - \frac{(\alpha+\rho)C_1}{k} - \mathbb{E}\left[(\gamma_i^*)^2\right]|V_{i,i}| - \frac{C_1^2}{k^2}\left\|[V]^i\right\|_1,$$

(2). By (2.8.57) in the above claim,

$$\mathbb{E}[\gamma_j^* \phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)] \le |v_j|\mathbb{E}[(\gamma_j^*)^2] + \frac{C_1^2}{k^2}(\|v\|_1 + \sigma),$$

(3). By (2.8.55), for $j \ne i$,

$$\mathbb{E}[\phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)] \le \mathbb{E}[|\sigma\gamma_i^* + \langle v, \gamma^*\rangle|] \le \frac{(\sigma + \|v\|_1)C_1}{k}.$$

Putting together, we can obtain

$$\begin{aligned}\mathbb{E}[\gamma_i^2] \ge \left(\Sigma_{i,i}^{-1} - |V_{i,i}|\right)^2 \mathbb{E}[(\gamma_i^*)^2] - \|[V]^i\|_2^2 \max_{j\in[k]} \mathbb{E}[(\gamma_j^*)^2] \\ - \left(\frac{C_1^2}{k^2}\|[V]^i\|_1(\|[V]^i\|_1 + 2\Sigma_{i,i}^{-1}) + \frac{2(\alpha+\rho)C_1}{k}\Sigma_{i,i}^{-1}\right).\end{aligned}$$

Second, we proceed to the upper bound. Similarly as the lower bound, by (2.8.55), we have

$$\begin{aligned}\mathbb{E}[\gamma_i^2] &\le \mathbb{E}\left[\left((|v_i| + \sigma)\gamma_i^* + \sum_{j\ne i}|v_j|\gamma_j^*\right)\phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)\right] \\ &= (|v_i| + \sigma)\mathbb{E}[\gamma_i^*\phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)] + \sum_{j\ne i}|v_j|\mathbb{E}[\gamma_j^*\phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)].\end{aligned}$$

For the first summand, same as in (2), by (2.8.57) in the above claim we get

$$\begin{aligned}\mathbb{E}[\gamma_i^*\phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)] &\le (\sigma + |v_i|)\mathbb{E}[(\gamma_i^*)^2] + \frac{C_1^2}{k^2}\|v\|_1, \\ \mathbb{E}[\gamma_j^*\phi_\alpha(\sigma\gamma_i^* + \langle v, \gamma^*\rangle + \xi_i)] &\le |v_j|\mathbb{E}[(\gamma_j^*)^2] + \frac{C_1^2}{k^2}(\|v\|_1 + \sigma).\end{aligned}$$

Therefore, we get

$$\mathbb{E}[\gamma_i^2] \le \left(\Sigma_{i,i}^{-1} + |V_{i,i}|\right)^2 \mathbb{E}[(\gamma_i^*)^2] + \|[V]^i\|_1(\|[V]^i\|_1 + 2\Sigma_{i,i}^{-1})\frac{C_1^2}{k^2} + \|[V]^i\|_2^2 \max_{j\in[k]} \mathbb{E}[(\gamma_j^*)^2].$$

which completes the proof. □

2.8.7.4 Main theorem

Theorem 71 (Main: Equilibration)**.** *If there exists an absolute constant $\mathcal{G}$ such that Assumption **(A1)**-**(A3)** and **(N1)** are satisfied with $l = 1/50$, $C_1^3 \le \mathcal{G}c_2^2 k$, $\max\left\{C_\nu, \|N^{(0)}\|_\infty\right\} \le \frac{\mathcal{G}c_2^4}{C_1^5 k\|(\beta^*)^\dagger\|_\infty}$, and additionally $\Sigma^{(0)} \le (1-\ell)\mathrm{Id}$, and $E \ge 0$ entry-wise, then there exist α, η, T, λ such that for sufficiently small $\epsilon > 0$ and sufficiently large $D = \mathrm{poly}(k, n, 1/\epsilon, 1/\delta)$ the following hold with probability at least $1-\delta$: Algorithm 9 outputs a solution $\beta = \beta^* D(\Sigma+E)+N$ where $\Sigma \ge (1-\ell)\mathrm{Id}$ is diagonal, $\|E\|_s \le \gamma\ell$ is off-diagonal, $\|N\|_\infty \le 2\|N^{(0)}\|_\infty$, and D is diagonal and satisfies*

$$\frac{\max_{i\in[k]} \frac{1}{D_{i,i}^2}\mathbb{E}[(\gamma_i^*)^2]}{\min_{j\in[k]} \frac{1}{D_{j,j}^2}\mathbb{E}[(\gamma_j^*)^2]} \le 2.$$

*If Assumption **(A1)**-**(A3)** and **(N2)** are satisfied with the same parameters except $\max\left\{C_\nu, \|N^{(0)}\|_\infty\right\} \le \min\left\{\sqrt{\frac{\mathcal{G}c_2^4}{C_1^5 k}}\frac{1}{\|(\beta^*)^\dagger\|_\infty}, \frac{\mathcal{G}c_2^2}{C_1^3\|(\beta^*)^\dagger\|_\infty}\right\}$, then the same guarantees hold.*

Proof of Theorem 71. The theorem follows from Lemma 77 (taking union bound over all the iterations and setting a proper δ), if the conditions are satisfied. So in the following, we first specify the parameters and then verify the conditions in Lemma 72 and Lemma 77.

Recall that $\ell = 1/50$. Define $u = 1+\ell$, $C = 3/2$, $B = 4$, $\kappa = 2$, $b = 3/4$, and let $\epsilon < 1/1000$.

Conditions in Lemma 72. For (2.8.36), we need to compute $r_i R_i$ and the the third term. Note that by the induction in Lemma 77, the M_j is an good approximation of $\mathbb{E}[(\gamma_j^*)^2]/(D_{j,j})^2$. Furthermore, when Lemma 72 is applied in Lemma 77, it is applied on the ground-truth matrix $(\beta^*)' = \beta^* D$ and $(\gamma_j^*)' = \gamma_j^*/D_{j,j}$, so M_j is a good approximation of $\mathbb{E}[((\gamma_j^*)')^2]$. Then

$$r_i R_i = \frac{3\mathbb{E}[((\gamma_i^*)')^2]}{5M_i} \le \frac{3}{5\left(\left(\frac{1}{u} - h_6\right)^2 - \kappa h_5^2 - \frac{1}{100}\right)}.$$

For the third term, first note that $C_1^3 \le \mathcal{G}c_2^2 k$, and thus $C_1^2 \le \mathcal{G}c_2 k$ by $C_1 > c_2$. Furthermore, $r_i = O(1/M_i) = O(k/c_2)$ for $i \in S$. Plugging in the parameters, we know that the third term is less than $1/1000$ when $\mathcal{G}$ is sufficiently small. Then (2.8.36) can be verified by plugging the parameters.

Similarly, for (2.8.37), we can compute $r_i R_i$ and let $\mathcal{G}$ small enough so that the second term is less than $1/1000$, and then verify the condition.

For (2.8.38) (2.8.39) and (2.8.40), we need to bound h_1 and h_2, which in turn relies on r and rR. Since for $i \in S$, $r_i = O(k/c_2)$, $r = O(k/c_2)$. Then similar to the argument as above, $h_1 < 2/10000$ when $\mathcal{G}$ is sufficiently small. when Lemma 72 is applied in Lemma 77, it is applied on the ground-truth matrix $(\beta^*)' = \beta^* D$ and $(\gamma_j^*)' = \gamma_j^*/D_{j,j}$. By the induction claims there, $\max_{j\in[k]} \mathbb{E}[((\gamma_j^*)')^2]$ differ from $\min_{j\in S} \mathbb{E}[((\gamma_j^*)')^2]$ by a factor of at most κ, so $rR \le \frac{3\kappa}{5}$. So the first term can be computed. The second term is less than $1/10000$ when $\mathcal{G}$ is small enough. Then h_2 can be computed. And the conditions can be verified.

Condition (2.8.41) is true since $\max\{C_\nu, \|N\|_\infty\} = O(\frac{c_2^2}{C_1^3\|(\beta^*)^\dagger\|_\infty})$. Condition (2.8.42) is true by setting $\epsilon' < U/8$ and by $U_a < U/8$ and $U = \|N\|_\infty \le O(\frac{c_2^2}{C_1^3\|(\beta^*)^\dagger\|_\infty})$. Similarly, condition (2.8.42) is true by setting $\epsilon' < U/8$ and by $U_n < U/8$ and $\|N\|_\infty$ is sufficiently small.

Conditions in Lemma 77. First, consider (2.8.53) and (2.8.54). As mentioned above, since $C_1^3 = O(c_2^2 k)$ and $C_1^2 = O(c_2 k)$, then h_3 and h_4 can be made sufficiently small to satisfy the conditions. (2.8.51) and (2.8.52) can be verified by plugging (2.8.53) and (2.8.54) and the assumption that $\lambda \ge \min_{i\in[k]} \mathbb{E}[(\gamma_i^*)^2]/\kappa$.

This completes the proof. □

2.8.8 Technical details: auxiliary lemmas for solving recurrences

The following lemmas are used when solving some of the recurrences in our analysis of the updates:

Lemma 79 (Coupling update rule)**.** *Let $\{a_t\}_{t=0}^\infty, \{b_t\}_{t=0}^\infty$ be sequences of non-negative numbers such that for fixed values $h \ge 0$, $\eta \in [0,1]$, $R > 4r > 0$:*

$$\begin{aligned} a_{t+1} &\le (1-\eta)a_t + \eta r b_t + \eta h \\ b_{t+1} &\le (1-\eta)b_t + \frac{\eta}{R} a_t + \eta h \end{aligned}$$

Then the following two properties holds:

1. $$\forall t \ge 0, a_t + b_t \le a_0 + b_0 + \frac{Rr + 2R + 1}{R - r} h$$

2. *For all $\epsilon > 0$, when $t \ge \ln \frac{a_0+b_0}{8\eta\epsilon}$, we have:*

$$a_t \le \frac{R(r+1)}{R-r} h + \epsilon, \quad b_t \le \frac{R+1}{R-r} h + \epsilon$$

Proof of Lemma 79. Observe that the update rule is equivalent to

$$\begin{aligned}\left(a_{t+1} - \frac{R(r+1)}{R-r}h\right) &\leq (1-\eta)\left(a_t - \frac{R(r+1)}{R-r}h\right) + \eta r\left(b_t - \frac{R+1}{R-r}h\right)\\ \left(b_{t+1} - \frac{R+1}{R-r}h\right) &\leq (1-\eta)\left(b_t - \frac{R+1}{R-r}h\right) + \frac{\eta}{R}\left(a_t - \frac{R(r+1)}{R-r}h\right)\end{aligned}$$

Therefore, define $c_t = a_t - \frac{R(r+1)}{R-r}h$ and $d_t = b_t - \frac{R+1}{R-r}h$, we can rewrite above as:

$$\begin{aligned}c_{t+1} &\leq (1-\eta)c_t + \eta r d_t\\ d_{t+1} &\leq (1-\eta)d_t + \frac{\eta}{R}c_t\end{aligned}$$

Since we just need to upper bound c_t, d_t. without lose of generality, we can assume that

$$\begin{aligned}c_{t+1} &= (1-\eta)c_t + \eta r d_t\\ d_{t+1} &= (1-\eta)d_t + \frac{\eta}{R}c_t\end{aligned}$$

Which implies that

$$\begin{aligned}\left(c_{t+1} + \sqrt{\frac{R}{r}}d_{t+1}\right) &= \left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)\left(c_t + \sqrt{\frac{R}{r}}d_t\right)\\ \left(c_{t+1} - \sqrt{\frac{R}{r}}d_{t+1}\right) &= \left(1-\eta-\eta\sqrt{\frac{r}{R}}\right)\left(c_t - \sqrt{\frac{R}{r}}d_t\right)\end{aligned}$$

Which can be simplified to

$$\begin{aligned}\left(c_t + \sqrt{\frac{R}{r}}d_t\right) &= \left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t\left(c_0 + \sqrt{\frac{R}{r}}d_0\right)\\ \left(c_t - \sqrt{\frac{R}{r}}d_t\right) &= \left(1-\eta-\eta\sqrt{\frac{r}{R}}\right)^t\left(c_0 - \sqrt{\frac{R}{r}}d_0\right)\end{aligned}$$

Therefore, we can solve

$$c_t = \frac{1}{2}\left[\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t + \left(1-\eta-\eta\sqrt{\frac{r}{R}}\right)^t\right]c_0 + \frac{1}{2}\sqrt{\frac{R}{r}}\left[\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t - \left(1-\eta-\eta\sqrt{\frac{r}{R}}\right)^t\right]d_0$$

$$d_t = \frac{1}{2}\sqrt{\frac{r}{R}}\left[\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t - \left(1-\eta-\eta\sqrt{\frac{r}{R}}\right)^t\right]c_0 + \frac{1}{2}\left[\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t + \left(1-\eta-\eta\sqrt{\frac{r}{R}}\right)^t\right]d_0$$

Observe that for every $t \geq 0$, $a \geq b \geq 0$, $a^t - b^t \leq (a-b)ta^{t-1}$

Which implies:

$$\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t-\left(1-\eta-\eta\sqrt{\frac{r}{R}}\right)^t\leq 2t\eta\sqrt{\frac{r}{R}}\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^{t-1}$$

Therefore, when $c_0, d_0 \geq 0$,

$$c_t\leq\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t c_0+t\eta\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^{t-1} d_0$$

Moreover,

$$d_t\leq\frac{r}{R}\eta\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t c_0+\left(1-\eta+\eta\sqrt{\frac{r}{R}}\right)^t d_0$$

Taking the optimal t, we obtain $c_t + d_t \leq c_0 + d_0$, which implies that

$$a_t+b_t\leq a_0+b_0+\frac{Rr+2R+1}{R-r}h$$

On the other hand, when $t \geq \ln\frac{c_0+d_0}{8\eta\epsilon}$, $c_t, d_t \leq \epsilon$, which implies that

$$a_t\leq\frac{R(r+1)}{R-r}h+\epsilon,\quad b_t\leq\frac{R+1}{R-r}h+\epsilon.$$

□

Lemma 80 (Simple coupling)**.** *Let $\{a_t\}_{t=0}^{\infty}, \{b_t\}_{t=0}^{\infty}$ be sequences of non-negative numbers such that for fixed values $h_1, h_2 \geq 0$, $\eta \in [0, 1]$, $r > 0$:*

$$\begin{aligned}a_{t+1} &\leq (1-\eta)a_t+\eta h_1\\ b_{t+1} &\leq (1-\eta)b_t+\eta s a_t+\eta h_2\end{aligned}$$

Then

$$\begin{aligned}a_t &\leq u_a:=\max\{a_0,h_1\},\\ b_t &\leq \max\{b_0,h_2+su_a\}.\end{aligned}$$

Proof. We have

$$\begin{aligned}(a_{t+1}-h_1) &\leq (1-\eta)(a_t-h_1)\\ (b_{t+1}-h_2) &\leq (1-\eta)(b_t-h_2)+\eta s a_t\end{aligned}$$

Solving the first one gives

$$a_t \leq u_a := \max\{a_0, h_1\}.$$

Then

$$(b_{t+1} - h_2) \quad \leq \quad (1-\eta)(b_t - h_2) + \eta s u_a$$

leads to

$$b_t \leq \max\{b_0, h_2 + s u_a\}.$$

□

Lemma 81 (Simple recursion)**.** *Let $\{a_t\}_{t=0}^{\infty}$ be a sequences of non-negative numbers such that for fixed values $h \geq 0$, $\eta \in [0, 1]$,*

$$a_{t+1} \leq (1-\eta)a_t + \eta h.$$

Then,

$$a_t \leq (1-\eta)^t a_0 + h,$$

and thus for $t \geq \frac{\ln(\epsilon/a_0)}{\ln(1-\eta)}$, we have

$$a_t \leq \epsilon + h.$$

Proof. We will prove by induction that $a_t \leq (1-\eta)^t a_0 + h$, which implies the statement of the lemma. The base case is trivial, so we proceed to the induction:

$$a_{t+1} \leq (1-\eta)\left((1-\eta)^t a_0 + h\right) + \eta h \leq (1-\eta)^{t+1} a_0 + h$$

as we need. □

Chapter 3

Provable guarantees for learning non-linear latent-variable models using the method of moments

In this chapter, we will present new results on provable guarantees for method-of-moments based techniques for learning latent-variable models.

First, in Section 3.1 we will review two general algorithmic paradigms for implementing the method of moments: tensor-decompositions (Subsection 3.1.1) and algorithms for non-negative matrix factorization for instances with separable structure (Subsection 3.1.2). We will review why these paradigms, while being quite generic, naturally require a certain *linear* structure in the latent-variable model we are considering.

Subsequently, we will provide new extensions of these paradigms to allow for *non-linearities* in the model. More concretely, we will focus on noisy-OR networks: a textbook example of a non-linear latent-variable model, often used to model the causal structure of diseases and symptoms. In Section 3.3, we provide tensor decomposition-based algorithms with provable guarantees for learning noisy-OR networks; this section is based on results in (Arora et al., 2017b). Next, in Section 3.4, we provide algorithms for non-linear analogues of non-negative matrix factorization, and show how to apply them to derive algorithms for learning noisy-OR networks.

3.1 Overview of the method of moments

In this section, we will briefly review tensor-decompositions and non-negative matrix factorization based algorithms for implementing the method of moments. Recalling from Chapter 1, the framework for the method of moments

consists of:

(1) Expressing the moments of x as functions of the model parameters:

$$\mathbb{E}[x^{\otimes k}] = f_\theta(x) \tag{3.1.1}$$

for some function f_θ depending on the model parameters.

(2) Calculating the empirical moments $\mathbb{E}_{x\sim\hat{p}}[x^{\otimes k}]$, and solving the system of equations (3.1.1) for θ.

The recent surge in interest in the method of moments has come from progress on provable techniques for structured matrix and tensor decomposition for performing task (2) of the above framework. These have been applied to simple latent variable models such as topic models (Arora et al., 2012c; 2013b), sparse coding models (Arora et al., 2015b; Ma et al., 2016), mixtures of Gaussians (Hsu and Kakade, 2013; Ge et al., 2015), hidden Markov models (Mossel and Roch, 2005), etc.

To keep the discussion simple, we will not introduce more of the latent-variable models above, and instead focus on surveying two approaches to implementing the method of moments in the case of topid models. The first approach will be based on *low-rank tensor decomposition* (Section 3.1.1), and the second on leveraging a structural assumption of *separability* of the topic-word matrix (Section 3.1.2). Similar techniques apply to many of the other models we mentioned.

Subsequently, we will describe what is different about noisy-OR networks, and how we deal with this difficulty.

3.1.1 Tensor decomposition techniques for learning topic models with Dirichlet priors

Let us proceed to reviewing how tensor decomposition algorithms are used for implementing the method of moments. The crucial lemma that brings learning topic models into the realm of tensor decomposition is the following structural result about the moments of the observables for LDA.

Lemma 82 ((Anandkumar et al., 2014))**.** *Keep the notation of topic models from Section 1.1, and assume a Dirichlet prior with parameters $\alpha_i, i \in [k]$, denoting $\alpha_0 = \sum_{i=1}^k \alpha_i$. Furthermore, let us denote*

$$M_1 = \mathbb{E}[e_{x_1}], \qquad , M_2 = \mathbb{E}[e_{x_1} \otimes e_{x_2}] - \frac{\alpha_0}{\alpha_0+1} M_1 \times M_1,$$

$$M_3 = \mathbb{E}[e_{x_1} \otimes e_{x_2} \otimes e_{x_3}] - \frac{\alpha_0}{\alpha_0+2} \left(\mathbb{E}[e_{x_1} \otimes e_{x_2} \otimes M_2] + \mathbb{E}[e_{x_1} \otimes M \otimes e_{x_2}] + \mathbb{E}[M_1 \otimes e_{x_1} \otimes e_{x_2}]\right) + \frac{2\alpha_0^2}{(\alpha_0+2)(\alpha_0+1)} M_1 \otimes M_1 \otimes M_1$$

Then,

$$M_2 = \sum_{i=1}^{k} \frac{\alpha_i}{\alpha_0(\alpha_0+1)} \beta_i^{\otimes 2}, \qquad M_3 = \sum_{i=1}^{k} \frac{2\alpha_i}{\alpha_0(\alpha_0+1)(\alpha_0+2)} \beta_i^{\otimes 3}$$

In other words, decomposing the tensor M_3 reveals the topic vectors β_i. A few remarks are in order:

- From the span of the matrix M_2 alone we can only hope to recover the span of the topic vectors β_i. Indeed, even if we assume that the number of components k is smaller than the dimension, the span of the top k components in the singular value decomposition of M_2 will coincide with the span of the means β_i, but that is all the information we can extract from M_2 – since low rank matrix decompositions are only unique up to unitary transformations.

- Finding a rank-1 decomposition, i.e. a decomposition of a tensor into the minimum number of rank-1 terms, can be NP-hard for arbitrary tensors (Hillar and Lim, 2013). For many interesting cases, it will be unique (in contrast to the case of matrices). [1] Moreover, the proof of the uniqueness will be "algorithmic": i.e. the algorithm for finding the decomposition will be the certificate that is unique.

Subsequently, the question is how to perform the **tensor decomposition** of the M_3 tensor. As mentioned, we cannot hope to do this without any assumptions on the tensor. The crucial insight was provided in (Anandkumar et al., 2014), who proved that the *tensor power method*, the tensor analogue of the well-known power method for calculating eigenvectors of matrices can be used, provided the components of the tensor are *orthogonal vectors*. More precisely, they show:

Lemma 83 (Tensor power method, (Anandkumar et al., 2014))**.** *Suppose* $M = \sum_{i=1}^{m} w_i a_i^{\otimes 3}$, *s.t.* $\langle a_i, a_j \rangle = 0, \forall i \neq j$. *Let* $v_0 \in \mathbb{R}^n$ *be an arbitrary vector, s.t. the set of numbers* $|w_i \langle a_i, v_0 \rangle|$ *has a unique largest element. Without loss of generality, let this number be* $|w_1 \langle a_1, v_0 \rangle|$, *and let the second largest be* $|w_2 \langle a_2, v_2 \rangle|$. *Furthermore, let* $v_t = \frac{M(I, v_{t-1}, v_{t-1})}{\|M(I, v_{t-1}, v_{t-1})\|}$. *Then,*

$$\|a_t - v_t\|^2 \leq \left(2w_1^2 \sum_{i \geq 2} \frac{1}{w_i^2} \right) \left| \frac{w_2 \langle a_2, v_0 \rangle}{w_1 \langle a_1, v_0 \rangle} \right|^{2^{t+1}}$$

The way this lemma can be used is apparent: if we sample v_0 at random, there is an inverse polynomial probability that $w_2 \langle a_2, v_0 \rangle \ll w_1 \langle a_1, v_0 \rangle$. More formally:

Lemma 84 (Initialization for tensor power method, (Anandkumar et al., 2014))**.** *Keep the setting of Lemma 83. For any constant* γ, *there is a constant* $\delta(\gamma)$, *s.t. with probability at least* $1 - \eta$ *over the choice of* $k^{1+\delta}$ *uniformly random unit vectors, at least one satisfies* $w_2 \langle a_2, v_0 \rangle \leq \gamma w_1 \langle a_1, v_0 \rangle$.

[1] Of course, on the flipside, tensor decompositions can be non-unique – in a manner that is not as easy to characterize as matrices.

Putting Lemmas 84 and 83 with appropriate choice of parameters, we get a polynomial-time algorithm for decomposing tensor with *orthogonal components*. But, in our scenario, while the tensor M_3 is low-rank, the components μ_i are *not* necessarily orthogonal. This brings us to the final ingredient: *whitening* the tensor.

Namely, suppose that a matrix W satisfies $W^\top M_2 W = I$. In our case of interest, $M_2 \succeq 0$, so one concrete way to choose W is for instance, $W = UD^{-1/2}$, where $M_2 UDV^\top$ is the singular value decomposition of M_2. Defining $\tilde{M}_3 = \sum_{i=1}^m w_i (W^\top \mu_i)^{\otimes 3}$, it's easy to see that:

(1) The rank-1 components $\tilde{\mu}_i = W^\top \mu_i$ of $\tilde{M}_3$are orthogonal.

(2) $\hat{M}_3 = M_3(W, W, W)$, where the notation $M_3(W, W, W)$ denotes the standard multilinear form evaluation of M_3:

$$M_3(W, W, W)_{i_1, j_1, k_1} = \sum_{i_2, j_2, k_2 \in [n]} (M_3)_{i_2, j_2, k_2} W_{i_2, i_1} W_{j_2, j_1} W_{k_2, k_1}$$

Finally, a crucial point we overlooked in the above discussion was robustness. Namely, even in the usual instances (e.g. mixture of Gaussians, topic models), due to finite sample errors, the empirical version of the tensor M_3 is only *approximately* low rank. The standard approach described above can only tolerate small amounts of noise – on the order of $1/d$ in injective norm. These kinds of guarantees will be too weak to get non-trivial guarantees for the setting of noisy-OR networks we will be interested in. However, recent work due to (Ma et al., 2016) substantially improved these error bounds. In particular, they prove:

Theorem 2 (Robust tensor decomposition, (Ma et al., 2016))**.** *Suppose the tensor $\tilde{M}_3$ satisfies*

$$\|\tilde{M}_3 - \sum_{i=1}^m a_i^{\otimes 3}\|_{\{1\},\{2,3\}} \leq \epsilon \tag{3.1.2}$$

where $a_i, i \in [m]$ are orthonormal vectors in $\mathbb{R}^n$, and for a tensor $T \in \mathbb{R}^{n_1 \times n_2 \times n_3}$, we define the $\|\cdot\|_{\{1\}\{2,3\}}$ norm as

$$\|T\|_{\{1\}\{2,3\}} := \sup_{\substack{x \in \mathbb{R}^{n_1}, y \in \mathbb{R}^{n_2 n_3} \\ \|x\|=1, \|y\|=1}} \sum_{\substack{i \in [n_1] \\ (j,k) \in [n_2] \times [n_3]}} x_i y_{jk} T_{ijk}$$

Then, there exists a polynomial-time algorithm that returns $\tilde{a}_i, i \in [m]$ in polynomial time that is $O(\epsilon)$-close to $a_i, i \in [m]$ in ℓ_2 norm up to permutation. [2]

The algorithm is based on finding good vectors to initialize the tensor power method by, which are gotten by looking at the top eigenvectors of random "unfoldings" of the tensor $\tilde{M}_3$. More details of this will follow in Section 3.3.4 – including a generalization of the original proof in (Ma et al., 2016) to handle substantially larger, but structured error.

[2] Precisely, here we meant that there exists a permutation π such that for every i, $\max_i \|\tilde{a}_{\pi(i)} - a_i\|, \leq O(\epsilon)$

3.1.2 Non-negative matrix factorization techniques for learning separable instances of topic models

The main benefit of the tensor decomposition-based techniques for implementing the method of moments surveyed in the previous section is that they are very generic; they suffer, however, from an obvious problem for practical implementation: just building the third-moment tensor requires n^3 time; implementing the tensor power method requires $\omega(k)$ restarts per component, raising the runtime to $\omega(n^3 k)$. The vocabulary size of a training set consisting of New York Times articles, for instance, is on the scale of $15k$ – which renders this approach less than appealing. Sensitivity to noise (both statistical and systemic, due to model mismatch) are often issues in practice as well, despite theoretical improvements to robusteness such as those in (Ma et al., 2016) we mentioned.

A parallel idea for implementing the method of moments was proposed by (Arora et al., 2012a; 2013a), using algorithms for solving structured instances of non-negative matrix factorization. The main idea is based on the following simple proposition:

Proposition 85 ((Arora et al., 2012a; 2013a))**.** *Using the notation of Section 2.2, the matrix* $\mathbb{E}[f \otimes f]$ *satisfies*

$$\mathbb{E}[f \otimes f] = \beta \mathbb{E}[\gamma\gamma^\top]\beta^\top$$

Furthermore, the matrices β *and* $\mathbb{E}[\gamma\gamma^\top]\beta^\top$ *have non-negative entries.*

As a consequence, we can hope to recover the matrix β by using *non-negative matrix factorization* techniques, which is the task of decomposing a given matrix Y as a product of two matrices A, X with non-negative entries, s.t. the number of rows of A (respectively columns of X) is minimized. This is called the *non-negative rank* of the matrix Y.

As is usual in problems involving matrix factorization with constraints on the factors, the problem is NP-hard in the worst case – and in fact, assuming the exponential time hypothesis (ETH) requires exponential time in the non-negative rank (Arora et al., 2012b). Perhaps even more seriously – the matrices A, X are in general not uniquely defined.

Fortunately, in (Arora et al., 2012a; 2013a), a natural assumption on the matrix β is identified, which allows the resulting non-negative matrix factorization problem to be provably solvable – with a very efficient algorithm even. The assumption they introduce is one of *anchor words*. A word is an *anchor word* for a topic, if it appears in that topic and no other. Under these assumptions, (Arora et al., 2013a) prove:

Theorem 86 ((Arora et al., 2013a))**.** *There is an algorithm that given* $\tilde{O}(\frac{k^3}{\epsilon^3 p^6})$ *documents, where* p *is the minimum probability of any anchor word, and runs in time* $\tilde{O}((n^2 + nk)\frac{1}{\epsilon^2})$ *, which recovers the topic-word matrix to entry-wise error* ϵ*.*

The algorithm proceeds by first identifying the anchor words for each of the topics – this is done by finding the

vertices of the polytope formed by the convex hull of the rows of the covariance matrix $\mathbb{E}[f \otimes f]$. In the earlier paper (Arora et al., 2012a), this was done by solving many linear programs, which was practically too expensive; in the later paper (Arora et al., 2013a), this is done by an efficient combinatorial algorithm. Subsequently to finding the anchor words, to recover the topics (Arora et al., 2012a) proceed by performing matrix inversion – which ends up being too unstable and non-robust to noise; (Arora et al., 2013a) notice this step has a probabilistic interpretation, which allows them to make it substantially more robust.

3.2 Beyond linearity: overview of the noisy-OR problem

Though we were focusing on topic models, other latent variable models for which tensor methods have been applied share one important characteristic: they are linear, in the sense that the moments (or the marginals, depending on the particular model) of the observed variables, conditioned on the latent variables depend linearly on the hidden variables. But many settings seem to call for nonlinearity in the model. For instance, Bayesian networks in many domains involve highly nonlinear operations on the latent variables, and could even have multiple layers. The study of neural networks also runs into nonlinear models such as restricted Boltzmann machines (RBM) (Smolensky, 1986; Hinton and Salakhutdinov, 2006).

We will consider here possibly the text-book example of a non-linear model: a noisy-OR network (Jordan et al., 1999). The canonical use of this model is to model the relationship between diseases and symptoms, as in the classical human-constructed tool for medical diagnosis called *Quick Medical Reference* (QMR-DT) by ((Miller et al., 1982), (Shwe and Cooper, 1991)) which captures relationships between 570 diseases and 4075 symptoms, with $45,470$ directed edges, and the W_{ij}'s are small integers. [3]

It is of course desirable to automatically learn the network from unlabeled data, without the use of human experts. With respect to prior work, previously there were no approaches that even heuristically work at the required problem size ($n = 4000$). (In contrast to the inference problem which has seen more work, including reasonable heuristic methods (Jordan et al., 1999)). (Halpern and Sontag, 2013; Jernite et al., 2013) have designed some algorithms for this problem, however their first paper (Halpern and Sontag, 2013) assumes the graph structure is given; the second paper (Jernite et al., 2013) requires the Bayes network to be quartet-learnable, which is a strong structural assumption on the network. Finally, we note that the problem of finding a "best-fit" Bayesian network according to popular metrics[4] has been shown to be NP-complete by (Chickering, 1996) even when all of the hidden variables are also observed.

We will consider two different sets of assumptions for this problem, and we will propose different (provable) algorithms for each one. In the first one, we will assume a kind of "genericity" on the weights – namely, the weights

[3] We thank Randolph Miller and Vanderbilt University for providing the current version of this network for our research.

[4] Researchers resort to these metrics as when the graph structure is unknown, multiple structures may have the same likelihood, so maximum likelihood is not appropriate.

will be random (from a natural distribution); more generally, we will be able to get guarantees under the assumption of a kind of conditioning on the *pointwise mutual matrix*: a quantity akin to the second moment matrix, which we define. The algorithm we use will be based on tensor decompositions, as described in Section 3.1.1. In the second one, we will assume a weakening of the anchor words assumption, as described in Section 3.1.2. In either case, the difficulty compared to prior work is to deal with the *non-linear* nature of the problem.

3.3 Beyond linearity I: provable algorithms for noisy-OR using tensor decomposition

We move to the first approach to the noisy-OR problem using tensor decompositions. On a high level, our algorithm will use a certain correlation measure called *pointwise mutual information* in place of the moments of the observable variables. Recalling the notation of Section 1.1, we note that the conditional probabilities of the symptoms in the noisy-OR model have the following succinct form:

$$\Pr[s_i = 0 \mid d] = \prod_{j=1}^{m} \exp(-W_{ij} d_j) = \exp(-\langle W^i, d \rangle)\,. \tag{3.3.1}$$

Thus, these conditional probabilities are *log-linear*: the logarithm of their value is linear in the model parameters. On a high level, our approach will use a particular measure of correlation involving the logarithms of the moments of the symptoms called *pointwise mutual information*. We will Taylor expand this quantity, which will turn the problematic exponential in (3.3.1) into an infinite sum, in which we will ignore all but the first two terms. This brings the problem into the realm of tensor decomposition described in Section 3.1.1, but with an important novel twist: the error from ignoring the higher-order terms is systemic (i.e. it doesn't go to zero as the number of samples increases) and too large to be handled by any conventional robust tensor decomposition analysis, including (Ma et al., 2016) described in that section. On the other hand, however, this error is highly structured, and we will be able to show that the methods described in Section 3.1.1 can nevertheless handle them.

3.3.1 Overview of the assumptions, algorithm and results

Assumptions : we make several assumptions about the weights W, some of which we verified the QMR-DT network, but the other assumptions are asymptotic in nature. Thus the cleanest description of our algorithm is in an average-case setting. First, we assume all priors for the diseases are equal, namely $\rho_i = \rho, \forall i \in [m]$, which should be thought of as small (In the QMR-DT application, ρ is like $O(1/m)$.) Next we assume that the ground truth $W \in \mathbb{R}^{n \times m}$ has entries

picked in iid fashion using the following random process:

$$W_{ij} = \begin{cases} 0, & \text{with probability } 1-p \\ \widetilde{W}_{ij}, & \text{with probability } p \end{cases}$$

where $\widetilde{W}_{ij}$'s are upper bounded by ν_u for some constant ν_u and are identically distributed according to a distribution $\mathcal{D}$ which satisfies that for some constant $\nu_l > 0$,

$$\mathop{\mathbb{E}}_{\widetilde{W}_{ij}\sim\mathcal{D}}\left[\exp(-\widetilde{W}_{ij}^2)\right] \leq 1-\nu_l\,. \tag{3.3.2}$$

The condition (3.3.2) intuitively requires that $\widetilde{W}_{ij}$ is bounded away from 0. We will assume that $p \leq 1/3$ and $\nu_u = O(1), \nu_l = \Omega(1)$. (Again, these are realistic for QMR-DT setting).

Under such conditions, informally, we will show the following claim:

Theorem 3 (Informally stated)**.** *There exists a polynomial time algorithm (Algorithm 17) that, given polynomially many samples from the noisy OR network described in the previous paragraph, recovers the weight matrix W with $\widetilde{O}(\rho\sqrt{pm})$ relative error in ℓ_2-norm in each column.*

Since we think of the prior of the diseases ρ as being on the order $O(1/m)$. This means that even if p is on the order of 1, our relative error bound equals to $O(1/\sqrt{m}) \ll 1$.

3.3.1.1 The algorithm in a nutshell

We will first provide an overview of the full algorithm, pointing out the main difficulties. We define a crucial quantity pointwise mutual information of two binary-valued random variables x and y is $PMI2(x,y) \triangleq \log \frac{\mathbb{E}[xy]}{\mathbb{E}[x]\,\mathbb{E}[y]}$. Note that it is positive iff $\mathbb{E}[x,y] > \mathbb{E}[x]\,\mathbb{E}[y]$ and thus is used as a measure of correlation in many fields. This concept can be extended in more than one way to a triple of boolean variables x, y, z and we use

$$PMI3(x,y,z) \triangleq \log \frac{\mathbb{E}[xy]\,\mathbb{E}[yz]\,\mathbb{E}[zx]}{\mathbb{E}[xyz]\,\mathbb{E}[x]\,\mathbb{E}[y]\,\mathbb{E}[z]}. \tag{3.3.3}$$

(We will sometimes shorten PMI3 and PMI2 to PMI when this causes no confusion.)

Our algorithm is given polynomially many samples from the model (recall, a sample contains only the observables: the symptoms that are or are not present in a particular patient). It starts by computing the following matrix $n \times n$ PMI and $n \times n \times n$ tensor PMIT, which tabulate the correlations among all pairs and triples of symptoms (specifically, the

indicator random variable for the symptom being absent):

$$\mathrm{PMI}_{ij} \triangleq PMI2(1-s_i, 1-s_j). \tag{3.3.4}$$

$$\mathrm{PMIT}_{i,j,k} \triangleq PMI3(1-s_i, 1-s_j, 1-s_j) \tag{3.3.5}$$

The next proposition makes the key observation that the above matrix and tensor are close to rank m, which we recall is much smaller than n. For convenience, we define $F, G \in \mathbb{R}^{n\times m}$ as

$$F \triangleq 1 - \exp(-W) \tag{3.3.6}$$

$$G \triangleq 1 - \exp(-2W)\,. \tag{3.3.7}$$

Proposition 87 (Informally stated)**.** *In the described setting,*

$$\mathrm{PMI} \approx \rho\left(FF^\top + \rho GG^\top\right) \quad = \quad \rho\sum_{k=1}^{m} F_k F_k^\top + \rho^2 \sum_{k=1}^{m} G_k G_k^\top \tag{3.3.8}$$

$$\tag{3.3.9}$$

$$\mathrm{PMIT} \approx \rho\Big(\underbrace{\sum_{k=1}^{m} F_k \otimes F_k \otimes F_k}_{:=S} + \underbrace{\rho\sum_{k=1}^{m} G_k \otimes G_k \otimes G_k}_{:=E}\Big)\,. \tag{3.3.10}$$

The proposition is proved later (with precise statement) in Section 3.3.5 by computing the moments by marginalization and using Taylor expansion to approximate the log of the moments, and ignoring terms ρ^3 and smaller. (Recall that ρ is the probability that a patient has a particular disease, which should be small, of the order of $O(1/n)$. The dependence of the final error upon ρ appears in Section 3.3.2.) Since the tensor PMIT can be estimated to arbitrary accuracy given enough samples, we have in some brought the problem somewhat closer to the linear tensor decomposition described in Section 3.1.1. Though this is the high level idea, the following difficulties have to be overcome.

Difficulty 1: Suppose in equation (3.3.10) we view the first summand S, which is rank m with components F_k's as the *signal* term. In all previous polynomial-time algorithms for tensor decomposition, the tensor is required to have the form $\sum_{k=1}^{m} F_k \otimes F_k \otimes F_k + noise$. To make our problem fit this template we could consider the second summand E as the "noise", especially since it is multiplied by $\rho \ll 1$ which tends to make E have smaller norm than S. But this is naive and incorrect, since E is a very structured matrix: it is more appropriate viewed as *systematic error*. (In particular this error doesn't go down in norm as the number of samples goes to infinity.) In order to do tensor decomposition in presence of such systematic error, we will need both a delicate error analysis and a very robust tensor decomposition algorithm. These will be outlined in Section 3.3.1.3.

Difficulty 2: To get our problem into a form suitable for tensor decomposition requires a *whitening* step, which uses

the robust estimate of the whitening matrix from the second moment matrix. In this case, the whitening matrix has to be extracted out of the PMI matrix, which itself suffers from a systematic error. This also is not handled in previous works, and requires a delicate control of the error. See Section 3.3.1.4 for more discussion.

Difficulty 3: There is another source of inexactness in equation (3.3.10), namely the approximation is only true for those entries with distinct indices — for example, the diagonal entry PMI_{ii} has completely different formula from that for PMI_{ij} when $i \neq j$. This will complicate the algorithm, as described in Subsections 3.3.1.3 and 3.3.1.4.

The next few Subsections sketch how to overcome these difficulties, and the details appear in the rest of the paper.

3.3.1.2 Recovering span of low-rank matrices in presence of systematic error

To illustrate the main ideas for dealing with the systematic error as described in Difficulty 1, we focus on a simpler task: approximately recovering the span of a low-rank matrix in the presence of systematic noise. The next section sketches an extension of this method to tensor decomposition with analogous systematic error.

In the classical setting of recovering a low-rank matrix in the presence of noise, there is an unknown $n \times n$ matrix S of rank m and we are given $S + E$ where E is an error matrix. The method to recover S is to compute the best rank-m approximation to $S + E$. The quality of this approximation was studied by Davis and Kahan (Davis and Kahan, 1970) and Wedin (Wedin, 1972), and many subsequent authors. The quality of the recovery depends upon the ratio $\|E\|/\sigma_m(S)$, where $\sigma_m(\cdot)$ denotes m-th largest singular value and $\|\cdot\|$ denotes the spectral norm. To make this familiar lemma fit our setting more exactly, we will phrase the problem as trying to recover a matrix S given noisy estimate $SS^\top + E$. Now one can only recover S up to rotation, and the following lemma describes the error in the Davis-Kahan recovery. It also plays a key role in the error analysis of the usual algorithm for tensor decomposition.

Lemma 88. *In the above setting, let $K, \widehat{K}$ the subspace of the top m eigenvectors of $SS^\top$ and $SS^\top + E$. Let ϵ be such that $\|E\| \leq \epsilon \cdot \sigma_m(SS^\top)$. Then $\left\|\text{Id}_K - \text{Id}_{\widehat{K}}\right\| \lesssim \epsilon$ where* Id *is the identity transformation on the subspace in question.*

The Lemma thus treats $\|E\|/\sigma_m(SS^\top)$ as the definition of *noise/signal* ratio. Before we generalize the definition and the algorithm to handle systematic error it is good to get some intuition, from looking at (3.3.8): $\text{PMI} \approx \rho(FF^T + \rho GG^T)$. Thinking of the first term as signal and the second as error, let's check how bad the noise/signal ratio defined in Davis-Kahan is. The "signal"is $\sigma_m(FF^\top)$, which is smaller than n since the trace of $FF^\top$ is of the order of mn in our probabilistic model for the weight matrix. The "noise" is the norm of $\rho GG^\top$, which is large since the G_k's are nonnegative vectors with entries of the order of 1, and therefore the quadratic form $\langle \frac{1}{\sqrt{n}}\mathbf{1}, \rho GG^\top \frac{1}{\sqrt{n}}\mathbf{1}\rangle$ can be as large as $\rho \sum_k \langle G_k, \frac{1}{\sqrt{n}}\mathbf{1}\rangle^2 \approx \rho mn$. Thus the Davis-Kahan noise/signal ratio is ρm, and so when $\rho m \ll 1$, it allows recovering the subspace of F with error $O(\rho m)$. Note that this is a vacuous bound since ρ needs to be at least $1/m$ so that the hidden variable d contains 1 non-zero entry in average. We'll argue that this error is too pessimistic and we can in fact drive the estimation error down to close to ρ.

Definition (spectral boundedness)**.** Let $n \geq m$. Let $E \in \mathbb{R}^{n\times n}$ be a symmetric matrix and $S \in \mathbb{R}^{n\times m}$. Then, we say E is τ-spectrally bounded by S if

$$E \preceq \tau(SS^\top + \sigma_m(SS^\top) \cdot \mathrm{Id}_n) \tag{3.3.11}$$

The smallest such τ is the "error/signal ratio" for this recovery problem.

This definition differs from Davis-Kahan's because of the $\tau SS^\top$ term on the right hand side of (3.3.11). This allows, for any unit vector x, the quadratic form value $x^T Ex$ to be as large as $\tau(x^T SS^\top x + \sigma_m(SS^\top))$. Thus for example the $\mathbf{1}$ vector no longer causes a large noise/signal ratio since both quadtratic forms $FF^\top$ and $GG^\top$ have large values on it.

This new error/signal ratio is no larger than the Davis-Kahan ratio, but can potentially be much smaller. Now we show how to do a better analysis of the Davis-Kahan recovery in terms of it. The proof of this theorem appears in Section 3.3.3.

Theorem 4 (matrix perturbation theorem for systematic error)**.** *Let $n \geq m$. Let $S \in \mathbb{R}^{n\times m}$ be of full rank. Suppose positive semidefinite matrix $E \in \mathbb{R}^{n\times n}$ is* ϵ-spectrally bounded *by $S \in \mathbb{R}^{n\times m}$ for $\epsilon \in (0, 1)$. Let $K, \widehat{K}$ the subspace of the top m eigenvectors of $SS^\top$ and $SS^\top + E$. Then,*

$$\left\|\mathrm{Id}_K - \mathrm{Id}_{\widehat{K}}\right\| \lesssim \epsilon\,.$$

Finally, we should consider what this new definition of noise/signal ratio achieves. The next proposition (whose proof appears in Section 3.3.6) shows that that under the generative model for W sketched earlier, $\tau = O(\log n)$. Therefore, $\sqrt{\rho}G$ is $\tilde{O}(\rho)$-bounded by F, and the recovery error of the subspace of F from $FF^\top + \rho GG^\top$ is $\tilde{O}(\rho)$ (instead of $O(\rho m)$ using Davis-Kahan).

Proposition 89. *Under the generative model for W, w.h.p, the matrix $G = 1 - \exp(-2W)$ is* τ-spectrally bounded *by $F = 1 - \exp(-W)$, with $\tau = \tilde{O}(1)$.*

Empirically, we can compute the τ value for the weight matrix W in the QMR-DT dataset (Shwe and Cooper, 1991), which is a textbook application of noisy OR network. For the QMR-DT dataset, τ is under 6. This implies that the recovery error of the subspace of F guaranteed by Theorem 4 is bounded by $O(\tau\rho) \approx \rho$, whereas the error bound by Davis-Kahan is $O(\rho m)$.

3.3.1.3 Tensor decomposition with systematic error

Now we extend the insight from the matrix case to tensor recovery under systematic error. In turns out condition (3.3.11) is also a good measure of error/signal for the tensor recovery problem of (3.3.10). Specifically, if G is τ-bounded by F, then we can recover the components F_k's from the PMIT with column-wise error $O(\rho\tau^{3/2}\sqrt{m})$. This requires a non-trivial algorithm (instead of SVD), and the additional gain is that we can recover F_k's individually, instead of only obtaining the subspace with the PMI matrix.

First we recall the prior state of the art for the error analysis of tensor decomposition with Davis-Kahan type bounds. The best error bounds involve measuring the magnitude of the noise matrix Z in a new way. For any $n_1 \times n_2 \times n_3$ tensor T, we define the $\|\cdot\|_{\{1\}\{2,3\}}$ norm as

$$\|T\|_{\{1\}\{2,3\}} := \sup_{\substack{x\in\mathbb{R}^{n_1}, y\in\mathbb{R}^{n_2 n_3} \\ \|x\|=1, \|y\|=1}} \sum_{\substack{i\in[n_1] \\ (j,k)\in[n_2]\times[n_3]}} x_i y_{jk} T_{ijk} \,. \tag{3.3.12}$$

Note that this norm is in fact the spectral norm of the flattening of the tensor (into a $n_1 \times n_2 n_3$ dimensional matrix). This norm is larger than the injective norm[5], but recently (Ma et al., 2016) shows that ϵ-error in this norm implies $O(\epsilon)$-error in the recovery guarantees of the components, whereas if one uses injective norm, the guarantees often pick up an dimension-dependent factor (Anandkumar et al., 2014). We define $\|\cdot\|_{\{2\}\{1,3\}}$ norm similarly. As is customary in tensor decomposition, the theorem is stated for tensors of a special form, where the components $\{u_i\}, \{v_i\}, \{w_i\}$ are orthonormal families of vectors. This can be ensured without loss of generality using a procedure called whitening that uses the 2nd moment matrix.

Theorem 5 (Extension of (**?**)Theorem 10.2]MSS16)**.** *There is a polynomial-time algorithm (Algorithm 11 later) which has the following guarantee. Suppose tensor T is of the form*

$$T = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i + Z$$

where $\{u_i\}, \{v_i\}, \{w_i\}$ are three collections of orthonormal vectors in $\mathbb{R}^d$, and $\|Z\|_{\{1\}\{2,3\}} \le \epsilon$, $\|Z\|_{\{2\}\{1,3\}} \le \epsilon$. Then, it returns $\{(\tilde{u}_i, \tilde{v}_i, \tilde{w}_i)\}$ in polynomial time that is $O(\epsilon)$-close to $\{(u_i, v_i, w_i)\}$ in ℓ_2 norm up to permutation. [6]

But in our setting the noise tensor has systematic error. An analog of Theorem 4 in this setting is complicated because even the whitening step is nontrivial. Recall also the inexactness in Proposition 87 due to the diagonal terms, which we earlier called *Difficulty 3*. We address this difficulty in the algorithm by setting up the problem using a

[5] The injective norm of the tensor T is defined as $\|T\|_{\{1\}\{2,3\}} := \sup_{\substack{x\in\mathbb{R}^{n_1}, y\in\mathbb{R}^{n_2} z\in\mathbb{R}^{n_3} \\ \|x\|=1, \|y\|=1, \|z\|=1}} \sum_{i\in[n_1], j\in[n_2], k\in[n_3]} x_i y_j z_k T_{ijk}$.

[6] Precisely, here we meant that there exists a permutation π such that for every i, $\max\{\|\tilde{u}_{\pi(i)} - u_i\|, \|\tilde{v}_{\pi(i)} - v_i\|, \|\tilde{w}_{\pi(i)} - w_i\|\} \le O(\epsilon)$

sub-tensor of the PMI tensor. Let S_a, S_b, S_c be a uniformly random equipartition of the set of indices $[n]$. Let

$$a_k = F_{k,S_a}, \quad b_k = F_{k,S_b}, \quad c_k = F_{k,S_c}, \tag{3.3.13}$$

where $F_{k,S}$ denotes the restriction of vector F_k to subset S. Moreover, let

$$\gamma_k = G_{k,S_a}, \quad \delta_k = G_{k,S_b}, \quad \theta_k = F_{k,S_c}. \tag{3.3.14}$$

Then, since the sub-tensor $\mathrm{PMIT}_{S_a,S_b,S_c}$ only contains entries with distinct indices, we can use Taylor expansion (see Lemma 95) to obtain that

$$\mathrm{PMIT}_{S_a,S_b,S_c} = \rho \sum_{k\in[m]} a_k \otimes b_k \otimes c_k + \rho^2 \sum_{k\in[m]} \gamma_k \otimes \delta_k \otimes \theta_k + \text{higher order terms}.$$

Here the second summand on the RHS corresponds to the second order term in the Taylor expansion. It turns out that the higher order terms are multiplied by ρ^3 and thus have negligible Frobenius norm, and therefore discussion below will focus on the first two summands.

For simplicity, let $T = \mathrm{PMIT}_{S_a,S_b,S_c}$. Our goal is to recover the components a_k, b_k, c_k from the approximate low-rank tensor T.

The first step is to whiten the components a_k's, b_k's and c_k's. Recall that $a_k = F_{k,S_a}$ is a non-negative vector. This implies the matrix $A = [a_1, \ldots, a_m]$ must have a significant contribution in the direction of the vector $\mathbf{1}$, and thus is far away from being well-conditioned. For the purpose of this section, we assume for simplicity that we can access the covariance matrix defined by the vector a_k's,

$$\bar{Q}_a := AA^\top = \sum_{k\in[m]} a_k a_k^\top. \tag{3.3.15}$$

Similarly we assume the access of $\bar{Q}_b$ and $\bar{Q}_c$ which are defined analogously. In Section 3.3.1.4 we discuss how to obtain approximately these three matrices.

Then, we can compute the whitened tensor by applying transformation $(\bar{Q}_a^+)^{1/2}, (\bar{Q}_b^+)^{1/2}, (\bar{Q}_c^+)^{1/2}$ along the three modes of the tensor T,

$$\begin{aligned}(\bar{Q}_a^+)^{1/2} \otimes (\bar{Q}_b^+)^{1/2} \otimes (\bar{Q}_c^+)^{1/2} \cdot T = \rho \sum_{k\in[m]} (\bar{Q}_a^+)^{1/2} a_k \otimes (\bar{Q}_b^+)^{1/2} b_k \otimes (\bar{Q}_c^+)^{1/2} c_k \\ \underbrace{+ \rho^2 \sum_{k\in[m]} (\bar{Q}_a^+)^{1/2} \gamma_k \otimes (\bar{Q}_b^+)^{1/2} \delta_k \otimes (\bar{Q}_c^+)^{1/2} \theta_k}_{:=Z} + \text{negligible terms}\end{aligned}$$

Now the first summand is a low rank orthogonal tensor, since $(\bar{Q}_a^+)^{1/2}a_k$'s are orthonormal vectors. However, the term Z is a systematic error and we use the following Lemma to control its $\|\cdot\|_{\{1\}\{2,3\}}$ norm.

Lemma 90. *Let $n \geq m$ and $A, B, C \in \mathbb{R}^{n\times m}$ be full rank matrices and let $\Gamma, \Delta, \Theta \in \mathbb{R}^{d\times \ell}$. Let $\gamma_i, \delta_i, \theta_i$ be the i-th column of Γ, Δ, Θ, respectively. Let $\bar{Q}_a = AA^\top, \bar{Q}_b = BB^\top, \bar{Q}_c = CC^\top$. Suppose $\Gamma\Gamma^\top$ (and $\Delta\Delta^\top$, $\Theta\Theta^\top$) is τ-spectrally bounded by A (and B, C respectively), then,*

$$\left\| \sum_{i\in[\ell]} (\bar{Q}_a^+)^{1/2}\gamma_i \otimes (\bar{Q}_b^+)^{1/2}\delta_i \otimes (\bar{Q}_c^+)^{1/2}\theta_i \right\|_{\{1\}\{2,3\}} \leq (2\tau)^{3/2}\,.$$

Lemma 90 shows that to give an upper bound on the $\|\cdot\|_{\{1\}\{2,3\}}$ norm of the error tensor Z, it suffices to show that the square of the components of the error, namely, $\Gamma\Gamma^\top, \Delta\Delta^\top, \Theta\Theta^\top$ are τ-spectrally bounded by the components of the signal A, B, C respectively. This will imply that $\|Z\|_{\{1\}\{2,3\}} \leq (2\tau)^{3/2}\rho^2$.

Recall that A and Γ are two sub-matrices of F and G. We have shown that $GG^\top$ is τ-spectrally bounded by F in Proposition 89. It follows straightforwardly that the random sub-matrices also have the same property.

Proposition 91. *In the setting of this section, under the generative model for W, w.h.p, we have that $\Gamma\Gamma^\top$ is τ-*spectrally bounded *by A with $\tau = O(\log n)$. The same is true for the other two modes.*

Using Proposition 91 and Lemma 90, we have that

$$\|Z\|_{\{1\}\{2,3\}} \lesssim \rho^2 \log^{3/2}(n)\,.$$

Then using Theorem 5 on the tensor $(\bar{Q}_a^+)^{1/2} \otimes (\bar{Q}_b^+)^{1/2} \otimes (\bar{Q}_c^+)^{1/2} \cdot T$, we can recover the components $(\bar{Q}_a^+)^{1/2}a_k$'s, $(\bar{Q}_b^+)^{1/2}b_k$'s, and $(\bar{Q}_c^+)^{1/2}c_k$'s. This will lead us to recover a_k,b_k and c_k, and finally to recover the weight matrix W.

3.3.1.4 Robust whitening

In the previous subsection, we assumed the access to $\bar{Q}_a, \bar{Q}_b, \bar{Q}_c$ (defined in (3.3.15)) which turns out to be highly non-trivial. A priori, using equation (3.3.8), noting that $A = [F_{1,S_a}, \ldots, F_{m,S_a}]$, we have

$$\mathrm{PMI}_{S_a,S_a}/\rho \approx \bar{Q}_a + \text{error}\,.$$

However, this approximation can be arbitrarily bad for the diagonal entries of PMI since equation (3.3.8) only works for entries with distinct indices. (Recall that this is why we divided the indices set into S_a, S_b, S_c and studied the asymmetric tensor in the previous subsection). Moreover, the diagonal of the matrix $\bar{Q}_a$ contributes to its spectrum significantly and therefore we cannot get meaningful bounds (in spectral norm) by ignoring the diagonal entries.

This issue turns out to arise in most of the previous tensor papers and the solution was to compute $AA^\top$ by using the asymmetric moments $AB^\top, BC^\top, CA^\top$,

$$AA^\top = (AB^\top)(CB^\top)^+(CA^\top)\,.$$

Typically $AB^\top, BC^\top, CA^\top$ can be estimated with arbitrarily small error (as number of samples go to infinity) and therefore the equation above leads to accurate estimate to $AA^\top$. However, in our case the errors in the estimate $\mathrm{PMI}_{S_a,S_b} \approx AB^\top$, $\mathrm{PMI}_{S_b,S_c} \approx BC^\top$, $\mathrm{PMI}_{S_c,S_a} \approx CA^\top$ are systematic. Therefore, we need to use a more delicate analysis to control how the error accumulates in the estimate,

$$\bar{Q}_a \approx \mathrm{PMI}_{S_a,S_b} \cdot \mathrm{PMI}^{-1}_{S_b,S_c} \cdot \mathrm{PMI}_{S_c,S_a}\,.$$

Here again, to get an accurate bound, we need to understand how the error in $\mathrm{PMI}_{S_a,S_b} - AB^\top$ behaves relatively compared with $AB^\top$ in a direction-by-direction basis. We generalized Definition to capture the asymmetric spectral boundedness of the error by the signal.

Definition (Asymmetric spectral boundedness)**.** Let $n \geq m$ and $B, C \in \mathbb{R}^{n\times m}$. We say a matrix $E \in \mathbb{R}^{n\times n}$ is ϵ-spectrally bounded by (B, C) if E can be written as:

$$E = B\Delta_1 C^\top + B\Delta_2^\top + \Delta_3 C^\top + \Delta_4\,. \tag{3.3.16}$$

Here $\Delta_1 \in \mathbb{R}^{m\times m}$, $\Delta_2, \Delta_3 \in \mathbb{R}^{n\times m}$ and $\Delta_4 \in \mathbb{R}^{n\times n}$ are matrices whose spectral norms are bounded by: $\|\Delta_1\| \leq \epsilon$, $\|\Delta_2\| \leq \epsilon\sigma_{min}(C)$, $\|\Delta_3\| \leq \epsilon\sigma_{min}(B)$ and $\|\Delta_4\| \leq \epsilon\sigma_{min}(B)\sigma_{min}(C)$.

Let K be the column subspace of B and H be the column subspace of C. Then we have $\Delta_1 = B^+E(C^\top)^+$, $\Delta_2 = B^+E\mathrm{Id}_{H^\perp}$, $\Delta_3 = \mathrm{Id}_{K^\perp}E(C^\top)^+$, $\Delta_4 = \mathrm{Id}_{K^\perp}E\mathrm{Id}_{H^\perp}$. Intuitively, they measure the relative relationship between E and B, C in different subspaces. For example, Δ_1 is the relative perturbation in the column subspace of K and row subspace of H. When $B = C$, this is equivalent to the definition in the symmetric setting (this will be clearer in the proof of Theorem 4).

Theorem 6 (Robust whitening theorem)**.** *Let $n \geq m$ and $A, B, C \in \mathbb{R}^{n\times m}$. Suppose $\Sigma_{ab}, \Sigma_{bc}, \Sigma_{ca} \in \mathbb{R}^{n\times n}$ are of the form,*

$$\Sigma_{ab} = AB^\top + E_{ab},\ \ \Sigma_{bc} = BC^\top + E_{bc},\ \ \text{and}\ \ \Sigma_{ca} = CA^\top + E_{ca}.$$

where E_{ab}, E_{bc}, E_{ca} are ϵ-spectrally bounded by (A, B), (B, C), (C, A) respectively. Then, the matrix matrix

$$Q_a = \Sigma_{ab}[\Sigma_{bc}^\top]_m^+\Sigma_{ca}$$

*is a good approximation of $AA^\top$ in the sense that $Q_a = \Sigma_{ab}[\Sigma_{bc}^\top]_m^+\Sigma_{ca} - AA^\top$ is $O(\epsilon)$-*spectrally bounded *by A. Here $[\Sigma]_m$ denotes the best rank-m approximation of Σ.*

The theorem is non-trivial even if the we have an absolute error assumption, that is, even if $\|E_{bc}\| \le \tau\sigma_{min}(B)\sigma_{min}(C)$, which is stronger condition than E_{bc} is τ-spectrally bounded by (B, C). Suppose we establish bounds on $\|\Sigma_{ab} - AB^\top\|$, $\|\Sigma_{bc}^{+\top} - (BC^\top)^+\|$ and $\|\Sigma_{ab} - AB^\top\|$ individually, and then putting them together in the obvious way to control the error $\Sigma_{ab}[\Sigma_{bc}^\top]_m^+\Sigma_{ca} - AB^\top(BC^\top)^+CA^\top$. Then the error will be too large for us. This is because standard matrix perturbation theory gives that $\|\Sigma_{bc}^{-\top} - (BC^\top)^{-1}\|$ can be bounded by $O\left(\|E_{bc}\|\|(BC^\top)^{-1}\|^2\right) \lesssim \epsilon/[\sigma_{min}(B)\sigma_{min}(C)]$, which is tight. Then we multiply the error with the norm of the rest of the two terms, the error will be roughly $\epsilon \cdot \frac{\sigma_{\max}(B)\sigma_{\max}(C)}{\sigma_{\min}(B)\sigma_{\min}(C)}$. That is, we will loss a condition number of B, C, which can be dimension dependent for our case.

The fix to this problem is to avoid bounding each term in $\Sigma_{ab}[\Sigma_{bc}^\top]_m^+\Sigma_{ca}$ individually. To do this, we will take the cancellation of these terms into account. Technically, we re-decompose the product $\Sigma_{ab}[\Sigma_{bc}^\top]_m^+\Sigma_{ca}$ into a new product of three matrices $(\Sigma_{ab}B^+)(B[\Sigma_{bc}^\top]_m^+C)(C^+\Sigma_{ca})$, and then bound the error in each of these terms instead. See Section 3.3.9 for details.

As a corollary, we conclude that the whitened vectors $(Q_a^+)^{1/2}a_i$'s are indeed approximately orthonormal.

Corollary 92. *In the setting of Theorem 6, we have that $(Q_a^+)^{1/2}A$ contains approximately orthonormal vectors as columns, in the sense that*

$$\|(Q_a^+)^{1/2}AA^\top(Q_a^+)^{1/2} - \mathrm{Id}\| \lesssim \epsilon\,.$$

Therefore we have found an approximate whitening matrix for A even though we do not have access to the diagonal entries.

3.3.2 Main Algorithms and Results

As sketched in Section 4.4, our main algorithm (Algorithm 17) uses tensor decomposition on the PMI tensor. In this section, we describe the different steps and how the fit together. Subsequently, all steps will be analyzed in separate sections.

Theorem 7 (Main theorem, random weight matrix, (Arora et al., 2017b))**.** *Suppose the true W is generated from the random model in Section 3.3.1 with $\rho pm \le c$ for some sufficiently small constant c. Then given $N = \mathrm{poly}(n, 1/p/, 1/\rho)$*

Algorithm 10 Learning Noisy-Or Networks via Decomposing PMI Tensor

Inputs: N samples generated from a noisy-or network, disease prior ρ
Outputs: Estimate of weight matrix $\widehat{W}$.

1. Compute the empirical PMI matrix and tensor $\widehat{\text{PMI}}$, $\widehat{\text{PMIT}}$ using equation (3.3.45).
2. Choose a random equipartition S_a, S_b, S_c of $[n]$.
3. Obtain approximate whitening matrices for $\widehat{\text{PMIT}}$ via Algorithm 13 for the partitioning S_a, S_b, S_c
4. Run robust tensor-decomposition Algorithm 12 to obtain vectors $\hat{a}_i, \hat{b}_i, \hat{c}_i, i \in [m]$
5. Let Y_i be the concatenation of the three vectors $\mathbf{1} - (\frac{1-\rho}{\rho})^{1/3}\hat{a}_i, \mathbf{1} - (\frac{1-\rho}{\rho})^{1/3}\hat{b}_i, \mathbf{1} - (\frac{1-\rho}{\rho})^{1/3}\hat{c}_i$. (Recall that $\hat{a}_i, \hat{b}_i, \hat{c}_i$ are of dimension $n/3$ each.)
6. **Return** $\widehat{W}$, where
$$\widehat{W}_{i,j} = \begin{cases} -\log((Y_i)_j), & \text{if } (Y_i)_j > \exp(-\nu_u) : \\ \exp(-\nu_u), & \text{otherwise} \end{cases}$$

number of examples, Algorithm 17 returns a weight matrix $\widehat{W}$ in polynomial time that satisfies

$$\forall i \in [m], \|\widehat{W}_i - W_i\|_2 \le \widetilde{O}(\eta\sqrt{pn}),$$

where $\eta = \tilde{O}\left(\sqrt{mp}\rho\right)$.

Note that the column ℓ_2 norm of W_i is on the order of $\sqrt{pn}$, and thus η can be thought of as the relative error in ℓ_2 norm. Note also that $\Pr[s_i = 0] = 1 - \Pr[s_i = 1] \approx 1 - pm\rho$, so $\rho pm = o(1)$ is necessary purely for sample complexity reasons. Finally, we can also state a result with a slightly weaker guarantee, but with only deterministic assumptions on the weight matrix W. Recall that $F = 1 - \exp(-W)$ and $G = 1 - \exp(-2W)$. We will also define third and fourth-order terms $H = 1 - \exp(-3W)$, $L = 1 - \exp(-4W)$.

We also define the incoherence of a matrix F. Roughly speaking, it says that the left singular vectors of F don't correlate with any of the natural basis vector much more than the average.

Definition (Incoherence:)**.** Let $F \in \mathbb{R}^{n\times m}$ have singular value decomposition $F = U\Sigma V^\top$. We say F is μ-incoherent if $\max_i \|U_i\| \le \sqrt{\mu m/n}$. where U_i is the i-th row of U.

We assume the weight matrix W satisfies the following deterministic assumptions,

1. $GG^\top, HH^\top, LL^\top$ is τ-spectrally bounded by F for $\tau \ge 1$.
2. F is μ-incoherent with $\mu \le \widetilde{O}(\sqrt{n/m})$.
3. If $\max_i \|F_i\|_0 \le pn$, with high probability over the choice of a subset $S_a, |S_a| = n/3$, $\sigma_{\min}(F_{S_a}) \gtrsim \sqrt{np}$ and $\rho pm \le c$ for some sufficiently small constant c.

Theorem 8 (Main theorem, deterministic weight matrix, (Arora et al., 2017b))**.** *Suppose the matrix W satisfies the*

conditions 1-3 above. Given polynomial number of samples, Algorithm 17 returns $\widehat{W}$ in polynomial time, s.t.

$$\forall i \in [m], \quad \|\widehat{W}_i - W_i\|_2 \leq \widetilde{O}(\eta \sqrt{np}) \,.$$

for $\eta = \sqrt{mp}\rho\tau^{3/2}$

Since the ℓ_2 norm of W_i is on the order of $\sqrt{np}$, the relative error in ℓ_2-norm is as most $\sqrt{m}\rho\tau^{3/2}$, which mirrors the randomized case above.

The proofs of Theorems uses the overall strategy of Section 4.4, and is deferred to Section 3.3.11. We give a high level outline that demonstrates how the proofs depends on the machinery built in the subsequent sections.

Both Theorem 7 and Theorem 8 are similarly proved – the only technical difference being how the third and higher order terms are bounded. (Because of generative model assumption, for Theorem 7 we can get a more precise control on them.) Hence, we will not distinguish between them in the coming overview.

Overall, we will follow the approach outlined in Section 4.4. Let us step through Algorithm 17 line by line:

1. The overall goal will be to recover the leading terms of the PMI tensor. Of course, we get samples only, so can merely get an empirical version of it. In Section 3.3.12, we show that the simple plug-in estimator does the job – and does so with polynomially many samples.

2. Recall *Difficulty 3* from Section 4.4 : the PMI tensor and matrix expression is only accurate on the off-diagonal entries. In order to address this, in Section 3.3.1.3 we passed to a sub-tensor of the original tensor by partitioning the symptoms into three disjoint sets, and considering the induced tensor by this partition.

3. In order to apply the robust tensor decomposition algorithm from Section 3.3.4, we need to first calculate whitening matrices. This is necessarily complicated by the fact that the diagonals of the PMI matrix are not accurate, as discussed in Section 3.3.1.4. Section 3.3.9 gives guarantees on the procedure for calculating the whitening matrices.

4. This is main component of the algorithm: the robust tensor decomposition machinery. In Section 3.3.4, the conditions and guarantees for the success of the algorithm are formalized. There, we deal with the difficulties layed out in Section 3.3.1.2 : namely that we have a substantial systematic error that we need to handle. (Both due to higher-order terms, and due to the missing diagonal entries)

5. This step, along with Step 6, is a post-processing step – which allows us to recover the weight matrix W after we have recovered the leading terms of the PMI tensor.

We also give a short quantitative sense of the guarantee of the algorithm. (The reader can find the full proof in Section 3.3.11.)

To get quantitative bounds, we will first need a handle on spectral properties of the random model: these are located in Section 3.3.6. As we mentioned above, the main driver of the algorithm is step 4, which uses our robust tensor decomposition machinery in Section 3.3.4. To apply the machinery, we first need to show that the second (and higher) order terms of the PMI tensor are spectrally bounded. This is done by applying Proposition 101, which roughly shows the higher-order terms are $O(\rho \log n)$-spectrally bounded by $\rho F F^\top$. The whitening matrices are calculated using machinery in Section 3.3.9. We can apply these tools since the random model gives rise to a $O(1)$-incoherent F matrix as shown in Lemma 3.3.7.

To get a final sense of what the guarantee is, the l_2 error which step 4 gives, via Theorem 10 roughly behaves like $\sqrt{\sigma_{\max}}\tau^{3/2}$, where $\sigma_{\max}$ is the spectral norm of the whitening matrices and τ is the spectral boundedness parameter. But, by Lemma 109 $\sigma_{\max}$ is approximately the spectral norm of $\rho F F^\top$ – which on the other hand by Lemma 98 is on the order of $mnp^2\rho$. Plugging in these values, we get the theorem statement.

3.3.3 Finding the Subspace under Heavy Perturbations

In this section, we show even if we perturb a matrix $SS^\top$ with an error whose spectral norm might be much larger than $\sigma_{min}(SS^\top)$, as long as E is spectrally bounded the top singular subspace of S is still preserved. We defer the proof of the asymmetric case (Theorem 6) to Section 3.3.9. We note that such type of perturbation bounds, often called relatively perturbation bounds, have been studied in (Ipsen, 1998; Li, 1998a;b; 1997). The results in these papers either require the that signal matrix is full rank, or the perturbation matrix has strong structure. We believe our results are new and the way that we phrase the bound makes the application to our problem convenient. We recall Theorem 4, which was originally stated in Section 4.4.

@title (matrix perturbation theorem for systematic error)**.** *Let $n \geq m$. Let $S \in \mathbb{R}^{n\times m}$ be of full rank. Suppose positive semidefinite matrix $E \in \mathbb{R}^{n\times n}$ is* ϵ-spectrally bounded *by $S \in \mathbb{R}^{n\times m}$ for $\epsilon \in (0,1)$. Let $K, \widehat{K}$ the subspace of the top m eigenvectors of $SS^\top$ and $SS^\top + E$. Then,*

$$\left\|\mathrm{Id}_K - \mathrm{Id}_{\widehat{K}}\right\| \lesssim \epsilon\,.$$

Proof. We can assume $\epsilon \leq 1/10$ since otherwise the statement is true (with a hidden constant 10). Since E is a positive semidefinite matrix, we write $E = RR^\top$ where $R = E^{1/2}$. Since A has full column rank, we can write $R = AS + B$ where $S \in \mathbb{R}^{m\times n}$ and the columns of B are in the subspace $K^\perp$. (Specifically, we can choose $S = A^+R$ and

$B = R - AA^+R = \mathrm{Id}_{K^\perp}B$.) By the definition of spectral boundedness, we have

$$\begin{aligned} BB^\top &= \mathrm{Id}_{K^\perp}RR^\top\mathrm{Id}_{K^\perp} \preceq \mathrm{Id}_{K^\perp}\epsilon\left(AA^\top + \sigma_m(AA^\top)\mathrm{Id}_n\right)\mathrm{Id}_{K^\perp}\,. \\ &= \epsilon\sigma_m(AA^\top)\mathrm{Id}_{K^\perp}\,. \end{aligned}$$

Therefore, we have that $\|B\|^2 \le \epsilon\sigma_{\min}(AA^\top)$. Moreover, we also have

$$\mathrm{Id}_K RR^\top \mathrm{Id}_K \preceq \epsilon AA^\top + \epsilon\sigma_{\min}\mathrm{Id}_K\,,$$

It follows that

$$AS\,S^\top A^\top \preceq 2\epsilon AA^\top\,.$$

which implies

$$\|S\,S^\top\| \le \epsilon.$$

Let $P = (\mathrm{Id}_m + S\,S^\top)^{1/2}$. Then we write $AA^\top + E$ as,

$$\begin{aligned} AA^\top + E = AA^\top + RR^\top &= AA^\top + (AS + B)(AS + B)^\top \\ &= A(\mathrm{Id} + S\,S^\top)A^\top + AS\,B^\top + BS^\top A^\top + BB^\top \\ &= (AP + BS^\top P^{-1})(AP + BS^\top P^{-1})^\top + BB^\top - BS^\top P^{-2}S\,B^\top \end{aligned} \tag{3.3.17}$$

Let $\widehat{A} = (AP + BS^\top P^{-1})$. Let K' be the column span of $\widehat{A}$. We first prove that $\widehat{K}$ is close to K'. Note that

$$\begin{aligned} \left\|BB^\top - BS^\top P^{-2}S\,B^\top\right\| &\lesssim \|B\|^2 + \|B\|^2\left\|S^\top P^{-2}S\right\| \lesssim \|B\|^2 && \text{(since } P = \mathrm{Id} + S\,S^\top \succeq S\,S^\top\text{)} \\ &\lesssim \epsilon\sigma_{\min}(AA^\top)\,. \end{aligned}$$

Moreover, we have $\sigma_{\min}(\widehat{A}\widehat{A}^\top) = \sigma_{\min}(\widehat{A})^2 = \left(\sigma_{\min}(AP) - \|BS^\top P^{-1}\|\right)^2 \ge (1 - O(\epsilon))\sigma_{\min}(A)^2$. Therefore, using Wedin's Theorem (Lemma 114) on equation (3.3.17), we have that

$$\|\mathrm{Id}_{\widehat{K}} - \mathrm{Id}_{K'}\| \lesssim \epsilon\,. \tag{3.3.18}$$

Next we show K' and K are also close. We have

$$\|\widehat{A} - AP\| \le \|BS^\top P^{-1}\| \le \epsilon \sqrt{\sigma_{\min}(A)^2} \qquad \text{(since } \|S\| \lesssim \sqrt{\epsilon}, \|B\| \lesssim \sqrt{\epsilon}\text{)}$$

Therefore, by Wedin's Theorem, K', as the span of top m left singular vectors of $\widehat{A}$, is close to the span of the top left singular vector of AP, namely, K

$$\|\mathrm{Id}_K - \mathrm{Id}_{K'}\| \lesssim \epsilon\,. \tag{3.3.19}$$

Therefore using equation (3.3.18) and (3.3.19) and triangle inequality, we complete the proof. □

3.3.4 Robust Tensor Decomposition with Systematic Error

In this section we discuss how to robustly find the tensor decomposition even in presence of systematic error. We first illustrate the main techniques in an easier setting of orthogonal tensor decomposition (Section 3.3.4.1), then we describe how it can be generalized to the general setting that we require for our algorithm (Section 3.3.4.2).

3.3.4.1 Warm-up: Approximate Orthogonal Tensor Decomposition

We start with decomposing an orthogonal tensor with systematic error. The algorithm we use here is a slightly more general version of an algorithm in (Ma et al., 2016).

Algorithm 11 Robust orthogonal tensor decomposition

Inputs: Tensor $T \in \mathbb{R}^{d\times d\times d}$, number $\delta, \epsilon \in (0,1)$.
Outputs: Set $S = \{(\tilde{a}_i, \tilde{b}_i, \tilde{c}_i)\}$

1. $S = \emptyset$
2. **For** $s = 1$ **to** $O(d^{1+\delta}\log d)$
3. Draw $g \sim \mathbb{N}(0, \mathrm{Id}_n)$, and compute $M = (\mathrm{Id}_n \otimes \mathrm{Id}_n \otimes g^\top)\cdot T$.[7]
4. Compute the top left and right singular vectors $u, v \in \mathbb{R}^d$ of M. Let $z = (u^\top \otimes v^\top \otimes \mathrm{Id}_n)\cdot T$.
5. If $(u^\top \otimes v^\top \otimes z^\top)\cdot T \ge 1 - \zeta$, where $\zeta = O(\epsilon)$, and u is $1/2$-far away from any of u_i's with $(u_i, v_i, w_i) \in S$, then add (u, v, w) to S.
6. **Return** S

Theorem 9 (Stronger version of Theorem 5)**.** *Suppose $\{u_i\}, \{v_i\}, \{w_i\}$ are three collection ϵ-approximate orthonormal*

[7] Recall that product of two tensor $(A\otimes B\otimes C)\cdot(E\otimes D\otimes F) = AE\otimes BD\otimes CF$

vectors. Suppose tensor T is of the form

$$T = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i + Z$$

with $\|Z\|_{\{2\}\{1,3\}} \le \tau$ and $\|Z\|_{\{1\}\{2,3\}} \le \tau$. Then, with probability at least 0.9, Algorithm 11 returns $S = \{(\tilde{u}_i, \tilde{v}_i, \tilde{w}_i)\}$ which is guaranteed to be $O((\tau+\epsilon)/\delta)$-close to $\{(u_i, v_i, w_i)\}$ in ℓ_2-norm up to permutation.

Proof Sketch of Theorem 9. The Theorem is a direct extension of (**?**)Theorem 10.2]MSS16 to asymmetric and approximate orthogonal case. We only provide a proof sketch here. We start by writing

$$M = \left(\mathrm{Id} \otimes \mathrm{Id} \otimes g^\top\right) \cdot T = \underbrace{\sum_{i=1}^{m} \langle g, w_i \rangle u_i v_i^\top}_{:=M_s} + \underbrace{(\mathrm{Id}_n \otimes \mathrm{Id}_n \otimes g^\top) \cdot Z}_{:=M_g} \tag{3.3.20}$$

Since $\|Z\|_{\{2\}\{1,3\}} \le \tau$ and $\|Z\|_{\{1\}\{2,3\}} \le \tau$, (**?**)Theorem 6.5]MSS16 implies that with probability at least $1 - d^2$ over the choice of g,

$$\left\|(\mathrm{Id}_n \otimes \mathrm{Id}_n \otimes g^\top) \cdot Z\right\| \le 2\sqrt{\log d} \cdot \tau$$

Let $t = 2\sqrt{\log d}$. We have that with probability $1/(d^{1+\delta}\log^{O(1)} d)$, $\langle g, w_1 \rangle \ge (1+\delta/3)t$ and $\langle g, w_j \rangle \le t$ for every $j \ne 1$. We condition on these events. Let $\bar{u}_i$ be a set of orthonormal vectors such that $E_u = [u_1, \ldots, u_m] - [\bar{u}_1, \ldots, \bar{u}_m]$ satisfies $\|E_u\| \le \epsilon$ (we can take $\bar{u}_i$'s to be the whitening of u_i's). Similarly define $\bar{v}_i$'s. Then we have that the term (defined in equation (3.3.20)) can be written as $\sum_i \langle g, w_1 \rangle \bar{u}_i \bar{v}_i + E'$ where $\|E\|' \lesssim \epsilon$. Let $\bar{M}_S = \sum_i \langle g, w_1 \rangle \bar{u}_i \bar{v}_i$. Then $\bar{M}_S$ has top singular value $\langle g, w_1 \rangle \ge (1+\delta/3)t$, and second singular value at most t. Moreover, the term $M_g + E'$ has spectral norm bounded by $O(\tau+\epsilon)$. Thus by Wedin's Theorem (Lemma 114), the top left and right singular vectors u, v of $M_S + M_g = \bar{M}_S + M_g + E'$ are $O((\tau+\epsilon)/\delta)$-close to $\bar{u}_1$ and $\bar{v}_1$ respectively. They are also $O((\tau+\epsilon)/\delta)$-close to u_1, v_1 since u_1 is close to $\bar{u}_1$. Moreover, we have $(u^\top \otimes v^\top \otimes \mathrm{Id}) \cdot T$ is $O(\tau/\delta)$-close to w_1.

Therefore, with probability $1/(d^{1+\delta}\log^{O(1)} d)$, each round of the for loop in Algorithm 11 will find u_1, v_1, w_1. Line 5 is used to verify if the resulting vectors are indeed good using the injective norm as a test. It can be shown that if the test is passed then (u, v, z) is close to one of the component. Therefore, after $d^{1+\delta}\log^{O(1)} d$ iterations, with high probability, we can find all of the components.

□

3.3.4.2 General tensor decomposition

In many previous works, general tensor decomposition is reduced to orthogonal tensor decomposition via a whitening procedure. However, here in our setting we cannot estimate the exact whitening matrix because of the systematic error.

Therefore we need a more robust version of approximate whitening matrix, which we define below:

Definition. Let $r \le d$. A collection of r vectors $\{a_1, \dots, a_r\}$ is ϵ-approximately orthonormal if the matrix A with a_i as columns satisfies

$$\left\|A^\top A - \mathrm{Id}\right\| \le \epsilon \tag{3.3.21}$$

Definition. Let $d \ge r$ and $A = [a_1, \dots, a_r] \in \mathbb{R}^{d\times r}$. A PSD matrix $Q \in \mathbb{R}^{d\times d}$ is an ϵ-approximate whitening matrix for A if $(Q^+)^{1/2}A$ is ϵ-approximately orthonormal .

Algorithm 12 Tensor decomposition with systematic error

Inputs: Tensor $T \in \mathbb{R}^{n_1\times n_2\times n_3}$ and ϵ-approximate whitening matrices $Q_a, Q_b, Q_c \in \mathbb{R}^{d\times d}$.
Outputs: $\{\hat{a}_i, \hat{b}_i, \hat{c}_i\}_{i\in[r]}$

1. Compute $\tilde{T} = (Q_a^+)^{1/2} \otimes (Q_b^+)^{1/2} \otimes (Q_c^+)^{1/2} \cdot T$
2. Run orthogonal tensor decomposition (Algorithm 11) with input $\tilde{T}$, and obtain $\{\check{a}_i, \check{b}_i, \check{c}_i\}$
3. **Return:** $\{Q_a^{1/2}\check{a}_i, Q_b^{1/2}\check{b}_i, Q_c^{1/2}\check{c}_i\}$

With this in mind, we can state the guarantee on the tensor decomposition algorithm (Algorithm 12).

Theorem 10. *Let $d \ge r$, and $A, B, C \in \mathbb{R}^{d\times r}$ be full rank matrices. Let $\Gamma, \Delta, \Theta \in \mathbb{R}^{d\times \ell}$. Let $a_i, b_i, c_i, \gamma_i, \delta_i, \theta_i$ be the columns of $A, B, C, \Gamma, \Delta, \Theta$ respectively. Suppose tensor T is of the form*

$$T = \sum_{i=1}^{r} a_i \otimes b_i \otimes c_i + \sum_{i=1}^{\ell} \gamma_i \otimes \delta_i \otimes \theta_i + E \tag{3.3.22}$$

Suppose matrices $Q_a \in \mathbb{R}^{d\times d}, Q_b \in \mathbb{R}^{d\times d}, Q_c \in \mathbb{R}^{d\times d}$ are ϵ-approximate whitening matrices for A, B, C, and suppose Γ, Δ, Θ are τ-spectrally bounded *by $Q_a^{1/2}, Q_b^{1/2}, Q_c^{1/2}$, respectively. Then, Algorithm 12 returns $\hat{a}_i, \hat{b}_i, \hat{c}_i$ that are $O(\eta)$-close to a_i, b_i, c_i in $\tilde{O}(d^{4+\delta})$ time with*

$$\eta \lesssim \max(\|Q_a\|, \|Q_b\|, \|Q_c\|)^{1/2} \cdot \left(\tau^{3/2} + \sigma^{-3/2}\|E\|_{\{1,2\}\{3\}} + \epsilon\right) \cdot 1/\delta$$

where $\sigma = \min(\sigma_{\min}(Q_a), \sigma_{\min}(Q_b), \sigma_{\min}(Q_c))$.

Note that in our model, the matrix E has very small spectral norm as it is the third order term in ρ (and $\rho = O(1/n)$). The spectral boundedness of Γ, Δ, Θ are discussed in Section 3.3.6. Therefore we can expect the RHS to be small.

In order to prove this theorem, we show after we apply whitening operation using the approximate whitening matrices, the tensor is still close to an orthogonal tensor. To do that, we need the following lemma which is a useful technical consequence of the condition (3.3.11).

Lemma 93. *Suppose F is τ-*spectrally bounded *by g. Then,*

$$\|G^\top (FF^\top)^+ G\| \le 2\tau\,. \tag{3.3.23}$$

Proof. Let K be the column span of F. Let $Q = FF^\top$. Multiplying $(Q^+)^{1/2}$ on both sides of equation (3.3.11), we obtain that

$$\begin{aligned}(Q^+)^{1/2} G G^\top (Q^+)^{1/2} &\le \tau(\mathrm{Id}_K + \sigma_m(Q) Q^+) \\ &\le \tau(\mathrm{Id}_K + \sigma_m(Q)\|Q^+\|\mathrm{Id}_K) \\ &\le 2\tau \mathrm{Id}_K\end{aligned}$$

It follows that $\|(Q^+)^{1/2} G\| \le \sqrt{2\tau}$, which in turns implies that $\|G^\top (FF^\top)^+ G\| = \|G^\top (Q^+)^{1/2} (Q^+)^{1/2} G\| \le 2\tau$. □

We also need to bound the $\{1,2\}\{3\}$ norm of the following systematic error tensor. This is important because we want to bound the spectral norm of the perturbation after the whitening operation.

Lemma 94 (Variant of (**?**)Theorem 6.1]MSS16)**.** *Let $\Gamma, \Delta, \Theta \in \mathbb{R}^{d\times \ell}$. Let $\gamma_i, \delta_i, \theta_i$ be the i-th column of Γ, Δ, Θ, respectively. Then,*

$$\left\| \sum_{i\in[\ell]} \gamma_i \otimes \delta_i \otimes \theta_i \right\|_{\{1,2\}\{3\}} \le \|\Gamma\| \cdot \|\Theta\| \cdot \|\Delta\|_{1\to 2} \le \|\Gamma\| \cdot \|\Theta\| \cdot \|\Delta\| \tag{3.3.24}$$

Proof of Lemma 94. Using the definition of $\|\cdot\|_{\{1,2\}\{3\}}$ we have that

$$\begin{aligned}\left\| \sum_{i\in[]} \gamma_i \otimes \delta_i \otimes \theta_i \right\|_{\{1,2\}\{3\}} &= \left\| \sum_{i\in[\ell]} (\gamma_i \otimes \delta_i)\theta_i^\top \right\| && (3.3.25)\\ &\le \left\| \sum_{i\in[\ell]} (\gamma_i \otimes \delta_i)(\gamma_i \otimes \delta_i) \right\|^{1/2} \left\| \sum_{i\in[\ell]} \theta_i \theta_i^\top \right\|^{1/2} && \text{(by Cauchy-Schwarz inequality)}\\ &= \left\| \sum_{i\in[\ell]} (\gamma_i \gamma_i^\top) \otimes (\delta_i \delta_i^\top) \right\|^{1/2} \|\Theta\| \end{aligned}$$

Next observe that we have that for any i, $\delta_i \delta_i^\top \preceq (\max \|\delta_i\|^2)\mathrm{Id}$ and therefore,

$$(\gamma_i \gamma_i^\top) \otimes (\delta_i \delta_i^\top) \preceq \gamma_i \gamma_i^\top \otimes (\max \|\delta_i\|^2)\mathrm{Id}\,. \tag{3.3.26}$$

It follows that

$$\left\|\sum_{i\in[r]}\gamma_i\otimes\delta_i\otimes\theta_i\right\|_{\{1,2\}\{3\}}\leq\left\|\sum_{i\in[r]}\gamma_i\gamma_i^\top\otimes(\max\|\delta_i\|^2)\mathrm{Id}\right\|^{1/2}\|\Theta\|$$
$$=\|\Gamma\|\cdot\|\Theta\|\cdot\|\Delta\|_{1\to2}\ .$$

□

With this in mind, we prove the main theorem:

Proof of Theorem 10. Let $\tilde{A}=(Q_a^+)^{1/2}A$, $\tilde{B}=(Q_b^+)^{1/2}B$, $\tilde{C}=(Q_c^+)^{1/2}C$. Moreover, let $\tilde{\Gamma}=(Q_a^+)^{1/2}\Gamma$ and define $\tilde{\Delta}$, $\tilde{\Theta}$ similarly. Let $\tilde{a}_1,\tilde{b}_i,\tilde{c}_i,\tilde{\gamma}_i,\tilde{\delta}_i,\tilde{\theta}_i$ be their columns. Then we have that $\tilde{T}$ as defined in Algorithm 12 satisfies

$$\tilde{T}=\sum_{i=1}^{r}\tilde{a}_i\otimes\tilde{b}_i\otimes\tilde{c}_i+\sum_{i=1}^{\ell}\tilde{\gamma}_i\otimes\tilde{\delta}_i\otimes\tilde{\theta}_i+\tilde{E}\tag{3.3.27}$$

where $\tilde{E}=(Q_a^+)^{1/2}\otimes(Q_b^+)^{1/2}\otimes(Q_c^+)^{1/2}\cdot E$. We will show that $\tilde{T}$ meets the condition of Theorem 5. Since Q_a is an ϵ-approximate whitening matrix of A, by Definition, $\tilde{A}=(Q_a^+)^{1/2}A$ is ϵ-approximately orthonormal . Similarly, $\tilde{B},\tilde{C}$ are ϵ-approximately orthonormal .

Γ is τ-spectrally bounded by Q_a, hence by Lemma 93, we have that $\|\tilde{\Gamma}\|\leq\sqrt{2\tau}$. Similarly, $\|\Theta\|,\|\Delta\|\leq\sqrt{2\tau}$. Applying Lemma 94, we have,

$$\left\|\sum_{i=1}^{\ell}\tilde{\gamma}_i\otimes\tilde{\delta}_i\otimes\tilde{\theta}_i\right\|_{\{1,2\}\{3\}}\leq(2\tau)^{3/2}\tag{3.3.28}$$

Moreover, we have $\|\tilde{E}\|_{\{1,2\}\{3\}}\leq\|(Q_a^+)^{1/2}\|\cdot\|(Q_c^+)^{1/2}\|\cdot\|(Q_c^+)^{1/2}\|\|E\|_{\{1,2\}\{3\}}\leq\sigma^{-3/2}\|E\|_{\{1,2\}\{3\}}$, where $\sigma=\min\{\sigma_{\min}(Q_a),\sigma_{\min}(Q_b),\sigma_{\min}(Q_c)\}$. Therefore, using Theorem 5 (with a_i,b_i,c_i there replaced by $\tilde{a}_i,\tilde{b}_i,\tilde{c}_i$, and Z there replaced by $\sum_{i=1}^{\ell}\tilde{\gamma}_i\otimes\tilde{\delta}_i\otimes\tilde{\theta}_i+\tilde{E}$), we have that a set of vectors $\{\breve{a}_i,\breve{b}_i,\breve{c}_i\}$ that are ϵ-close to $\{\tilde{a}_i,\tilde{b}_i,\tilde{c}_i\}$ with $\epsilon=(2\tau)^{3/2}+\sigma^{-3/2}\|\tilde{E}\|_{\{1,2\}\{3\}}$. Therefore, we obtain that $\|a_i-Q^{1/2}\breve{a}_i\|\leq\|Q_a\|^{1/2}\epsilon$. Similarly we can control the error for b_i and c_i and complete the proof. □

3.3.5 Formal expression for the PMI tensor

In this section we formally derive the expressions for the PMI tensors and matrices, which we only informally did in Section 4.4.

As a notational convenience for $l\in\mathbb{N}$, we will denote by $\tilde{P}_l$ the matrix which has as columns the vectors $1-$

$\exp(-lW_k), k \in [m]$. Furthermore, for a subset $S_a \subseteq [n]$, we will introduce the notation

$$P_{l,S_a} = \sum_{k\in[m]} \left((\tilde{P}_l)_{k,S_a}\right)\left((\tilde{P}_l)_{k,S_a}\right)^\top = \sum_{k\in[m]} (1-\exp(-lW_k)_{S_a})\,(1-\exp(-lW_k)_{S_a})^\top$$

These matrices will appear naturally in the expressions for the higher-order terms in the Taylor expansion for the PMI matrix and tensor.

We first compute the formally the moments of the noisy-or model.

Lemma 95. *We have*

$$\begin{aligned}
\log \Pr[s_i = 0] &= \sum_{k\in[m]} \log\left(1-\rho(1-\exp(W_{ik}))\right) \\
\forall i \neq j \log \Pr\left[s_i = 0 \wedge s_j = 0\right] &= \sum_{k\in[m]} \log\left(1-\rho(1-\exp(W_{ik}+W_{jk}))\right) \\
\forall \text{ distinct } i,j,k \in [n], \log \Pr\left[s_i = 0 \wedge s_j = 0 \wedge s_\ell = 0\right] &= \sum_{k\in[m]} \log\left(1-\rho(1-\exp(W_{ik}+W_{jk}+W_{\ell k}))\right)
\end{aligned}$$

Proof of Lemma 95. We only give the proof for the second equation. The rest can be shown analogously.

$$\begin{aligned}
\log \Pr\left[s_i = 0 \wedge s_j = 0\right] &= \mathbb{E}\left[\Pr[s_i = 0|d] \cdot \Pr[s_j = 0|d]\right] = \mathbb{E}\left[\exp(-(W_i+W_j)^\top d)\right] \\
&= \prod_{k\in[m]} \mathbb{E}\left[\exp(-(W_{ik}+W_{jk})d_k))\right] \\
&= \prod_{k\in[m]} \left(1-\rho(1-\exp(-(W_{ik}+W_{jk})))\right).
\end{aligned}$$

□

With this in mind, we give the expression for the PMI tensor along with all the higher-order terms.

Proposition 96. *For any equipartition S_a, S_b, S_c of $[n]$, the restriction of the PMI tensor $\mathrm{PMIT}_{S_a,S_b,S_c}$ satisfies, for any $L \geq 2$,*

$$\mathrm{PMIT}_{S_a,S_b,S_c} = \frac{\rho}{1-\rho} \sum_{k\in[m]} F_{k,S_a} \otimes F_{k,S_b} \otimes F_{k,S_c} + \sum_{l=2}^{L} (-1)^{l+1} \left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right) \sum_{k\in[m]} (\tilde{P}_l)_{k,S_a} \otimes (\tilde{P}_l)_{k,S_b} \otimes (\tilde{P}_l)_{S_c} + E_L \tag{3.3.29}$$

where

$$\|E_L\|_{\{1,2\},\{3\}} \leq \frac{(mn)^3}{L} \frac{\left(\frac{\rho}{1-\rho}\right)^L}{1-\left(\frac{\rho}{1-\rho}\right)^L}$$

Proof. The proof will proceed by Taylor expanding the log terms. Towards that, using Lemma 95, we have :

$$\mathrm{PMIT}_{ijl} = \sum_{k\in[m]} \log \frac{\left(1-\rho(1-\exp(-W_{ik}-W_{jk}))\right)(1-\rho(1-\exp(-W_{ik}-W_{lk})))\left(1-\rho(1-\exp(-W_{jk}-W_{lk}))\right)}{\left(1-\rho(1-\exp(-W_{ik}-W_{jk}-W_{lk}))\right)(1-\rho(1-\exp(-W_{ik})))\left(1-\rho(1-\exp(-W_{jk}))\right)(1-\rho(1-\exp(-W_{lk})))}$$

By the Taylor expansion of $\log(1-x)$, we get that

$$\begin{aligned}\mathrm{PMI}_{ijl} = &-\sum_{t=1}^{\infty}\frac{1}{t}\sum_{k\in[m]}\rho^t(\left(\left(1-\exp(-W_{ik}-W_{jk})\right)\right)^t + ((1-\exp(-W_{ik}-W_{lk})))^t + \left(\left(1-\exp(-W_{jk}-W_{lk})\right)\right)^t - \\ &(1-\exp(-W_{ik}))^t - \left(1-\exp(-W_{jk})\right)^t - (1-\exp(-W_{lk}))^t - \left(1-\exp(-W_{ik}-W_{jk}-W_{lk})\right)^t)\end{aligned}$$

Furthermore, note that

$$\begin{aligned}&\left(\left(1-\exp(-W_{ik}-W_{jk})\right)\right)^t + ((1-\exp(-W_{ik}-W_{lk})))^t + \left(\left(1-\exp(-W_{jk}-W_{lk})\right)\right)^t - \\ &(1-\exp(-W_{ik}))^t - \left(1-\exp(-W_{jk})\right)^t - (1-\exp(-W_{lk}))^t - \left(1-\exp(-W_{ik}-W_{jk}-W_{lk})\right)^t = \\ &\qquad\qquad \sum_{l=1}^{t}\binom{t}{l}(-1)^l\left(1-\exp(-lW_{ik})\right)\left(1-\exp(-lW_{jk})\right)\left(1-\exp(-lW_{lk})\right)\end{aligned}$$

by simple regrouping of the terms. By exchanging l and t, we get

$$\begin{aligned}\mathrm{PMIT}_{S_a,S_b,S_c} &= \sum_{l=1}^{\infty}\sum_{t\geq l}(-1)^{l+1}\left(\rho^t\frac{1}{t}\binom{t}{l}\right)\sum_{k\in[m]}(1-\exp(-lW_k))_{S_a}\otimes(1-\exp(-lW_k))_{S_b}\otimes(1-\exp(-lW_k))_{S_c} \\ &= \sum_{l=1}^{\infty}(-1)^{l+1}\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)\sum_{k\in[m]}(1-\exp(-lW_k))_{S_a}\otimes(1-\exp(-lW_k))_{S_b}\otimes(1-\exp(-lW_k))_{S_c} \qquad (3.3.30)\end{aligned}$$

where the last equality holds by noting that

$$\sum_{t\geq l}\rho^t\frac{1}{t}\binom{t}{l} = \frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l$$

The term corresponding to $t=1$ is easily seen to be

$$\frac{\rho}{1-\rho}\sum_{k\in[m]}F_{k,S_a}\otimes F_{k,S_b}\otimes F_{k,S_c}$$

therefore we to show the statement of the lemma, we only need bound the contribution of the terms with ł $\geq L$.

Toward that, note that $\forall l,k\|1-\exp(-lW_k)\|\leq n$. Hence, we have by Lemma 94,

$$\left\| \sum_{k=1}^{m} (1 - \exp(-lW_k))_{S_a} \otimes (1 - \exp(-lW_k))_{S_b} \otimes (1 - \exp(-lW_k))_{S_c} \right\|_{\{1,2\},\{3\}} \leq (mn)^3$$

Therefore, subadditivity of the $\{12\},\{3\}$ norm gives

$$\left\| \sum_{l=L}^{\infty} (-1)^{l+1} \left(\frac{\left(\frac{\rho}{1-\rho}\right)^l}{l} \right) \sum_{k\in[m]} (1 - \exp(-lW_k))_{S_a} \otimes (1 - \exp(-lW_k))_{S_b} \otimes (1 - \exp(-lW_k))_{S_c} \right\|_{\{1,2\},\{3\}}$$
$$\leq (mn)^3 \sum_{l=L}^{\infty} \left(\frac{\left(\frac{\rho}{1-\rho}\right)^l}{l} \right) \leq \frac{(mn)^3}{L} \sum_{l=L}^{\infty} \left(\frac{\rho}{1-\rho} \right)^l = \frac{(mn)^3}{L} \frac{\left(\frac{\rho}{1-\rho}\right)^L}{1 - \left(\frac{\rho}{1-\rho}\right)^L}$$

which gives us what we need. □

A completely analogous proof gives a similar expression for the PMI matrix:

Proposition 97. *For any subsets S_a, S_b of $[n]$, s.t. $S_a \cap S_b = \emptyset$, the restriction of the PMI matrix PMI_{S_a,S_b} satisfies, for any $L \geq 2$,*

$$\mathrm{PMI}_{S_a,S_b} = \frac{\rho}{1-\rho} \sum_{k\in[m]} F_{k,S_a} F_{k,S_b}^\top + \sum_{l=2}^{L} (-1)^{l+1} \left(\frac{1}{l} \left(\frac{\rho}{1-\rho} \right)^l \right) \sum_{k\in[m]} (\tilde{P}_l)_{k,S_a} ((\tilde{P}_l)_{k,S_b})^\top + E_L \tag{3.3.31}$$

where

$$\|E_L\|_{\{1,2\},\{3\}} \leq \frac{(mn)^2}{L} \frac{\left(\frac{\rho}{1-\rho}\right)^L}{1 - \left(\frac{\rho}{1-\rho}\right)^L}$$

3.3.6 Spectral properties of the random model

The goal of this section is to prove that the random model specified in Section **??** satisfies the incoherence property on the weight matrix and the spectral boundedness property of the PMI tensor. (Recall, the former is required for the whitening algorithm, and the later for the tensor decomposition algorithm.)

Before delving into the proofs, we will need a few simple bounds on the singular values of P_l.

Lemma 98. *Let $S_a \subseteq [n]$, s.t. $|S_a| = \Omega(n)$. With probability $1 - \exp(-\log^2 n)$ over the choice of W, and for all $l = O(poly(n))$,*

$$\sigma_{\min}(P_{l,S_a}) \gtrsim np$$

and

$$\sigma_{\max}(P_{l,S_a}) \lesssim mnp^2$$

Proof. Let us proceed to the lower bound first.

If we denote by L the matrix which has as columns $(\tilde{P}_l)_{k,S_a}$, $k \in [m]$, it's clear that $P_{l,S_a} = LL^\top$. Since

$$\sigma_{\min}(LL^\top) = \sigma_{\min}(L^\top L)$$

we will proceed to bound the smallest eigenvalue of $L^\top L$.

Note that

$$L^\top L = \sum_{k \in S_a} (1 - \exp(-lW^k))(1 - \exp(-lW^k))^\top$$

Since the matrices $(1-\exp(-lW^k))(1-\exp(-lW^k))^\top$ are independent, the bound will follow from a matrix Bernstein bound. Denoting

$$Q = \mathbb{E}\left[(1 - \exp(-lW^k))(1 - \exp(-lW^k))^\top\right]$$

by a simple calculation we have

$$Q = p^2\mathbb{E}\left[1 - \exp(-l\tilde{W})\right]^2 \mathbf{1}\mathbf{1}^\top + \left(p\mathbb{E}\left[(1 - \exp(-l\tilde{W}))^2\right] - p^2\mathbb{E}\left[1 - \exp(-l\tilde{w})\right]^2\right)\mathrm{Id}_m \tag{3.3.32}$$

where $\mathbf{1}$ is the all-ones vector of size m, and Id_m is the identity of the same size. Furthermore, $\tilde{W}$ is a random variable following the distribution $\mathcal{D}$ of all the $\tilde{W}_{i,j}$.

Note that (3.3.2) together with the assumption $\nu = \Omega(1)$ gives $\sigma_{\min}(Q) = \Omega(p)$

Let $Z^i = Q^{-1/2}(1 - \exp(-lW_i))$. Then we have that $\mathbb{E}\left[\sum_{i \in S_a} Z^i (Z^i)^\top\right] = |S_a| \cdot \mathrm{Id}_m$, and with high probability, $\|Z^i\|^2 \le 1/\sigma_{\min}(Q) \cdot \|F^i\|^2 \lesssim m$ and it's a sub-exponential random variable. Moreover,

$$r^2 = \left\|\sum_i \mathbb{E}\left[\left(\sum Z^i (Z^i)^\top\right)^2\right]\right\| \lesssim m \left\|\mathbb{E}\left[\sum_i Z^i (Z^i)^\top\right]\right\| \le mn\,.$$

Therefore, by Bernstein inequality we have that w.h.p,

$$\left\|\sum_{i \in S_a} Z^i (Z^i)^\top - n\mathrm{Id}_m\right\| \lesssim \sqrt{r^2 \log n} + \max \|Z^i\|^2 \log n = \sqrt{mn \log n} + m \log n\,.$$

It follows that

$$\sum_{i \in S_a} Z^i (Z^i)^\top \ge \left(n - O\left(\sqrt{mn \log n}\right)\right)\mathrm{Id}_m \ge n\mathrm{Id}_m\,.$$

which in turn implies that $P_{l,S_a} \ge nQ$. But this immediately implies $\sigma_{\min}(P_{l,S_a}) \gtrsim np$ with high probability. Union bounding over all l, we get the first part of the lemma.

The upper bound will be proven by a Chernoff bound. Note that the matrices

$$(1-\exp(-lW_k))_{S_a}(1-\exp(-lW_k))_{S_a}^\top, k \in [m]$$

are independent. Furthermore, $\|(1-\exp(-lW_k))_{S_a}(1-\exp(-lW_k))_{S_a}^\top\|^2 \leq pn$ with high probability, and the variable $\|(1-\exp(-lW_k))_{S_a}(1-\exp(-lW_k))_{S_a}^\top\|^2$ is sub-exponential. Finally,

$$\begin{aligned} r^2 &= \left\| \mathbb{E}\left[\sum_{k=1}^{m} ((1-\exp(-lW_k))_{S_a}(1-\exp(-lW_k))_{S_a}^\top)^2 \right] \right\| \\ &\leq \sum_{k=1}^{m} \left\| \mathbb{E}\left[((1-\exp(-lW_k)_{S_a})(1-\exp(-lW_k))_{S_a}^\top)^2 \right] \right\| \\ &\leq m \left\| 1-\exp(-lW_k)_{S_a} \right\|^2 \mathbb{E}\left[((1-\exp(-lW_k))_{S_a}(1-\exp(-lW_k))_{S_a}^\top) \right] \leq mn^2p^2 \end{aligned}$$

Similarly as in the lower bound,

$$\mathbb{E}[P_{l,S_a}] = p^2\mathbb{E}\left[1-\exp(-l\tilde{W})\right]^2 \mathbf{1}\mathbf{1}^\top + \left(p\mathbb{E}\left[(1-\exp(-l\tilde{W}))^2\right] - p^2\mathbb{E}\left[1-\exp(-l\tilde{w})\right]^2\right)\mathrm{Id}_{|S_a|}$$

where $\mathbf{1}$ is the all-ones vector of size $|S_a|$, and $\mathrm{Id}_{|S_a|}$ is the identity of the same size. Again, $\tilde{W}$ is a random variable following the distribution $\mathcal{D}$ of all the $\tilde{W}_{i,j}$. This immediately gives

$$P_l \leq \mathbb{E}[P_l] + r\log n\mathrm{Id}_{|S_a|} \leq mnp^2 + \sqrt{mn^2p^2}\log n\mathrm{Id}_{|S_a|} \leq O(mnp^2)\mathrm{Id}_{|S_a|}$$

A union bound over all values of l gives the statement of the Lemma.

□

3.3.7 Incoherence of matrix F

First, we proceed to show the incoherence property on the weight matrix.

Lemma 99. *Suppose n is a multiple of 3. Let $F = U\Sigma V$ be the singular value decomposition of F. Let S_a, S_b, S_c be a uniformly random equipartition of the rows of $[n]$. Suppose F is μ-incoherent with $\mu \leq n/(m\log n)$. Then, with high probability over the choice of and S_a, S_b, S_c, we have for every $i \in \{a,b,c\}$,*

$$\left\| \left(U^{S_i}\right)^\top U^{S_i} - \frac{1}{3}\mathrm{Id} \right\| \lesssim \sqrt{\frac{\mu m}{n}\log n}\,.$$

Proof. Let $S = S_a$. Then, since $U^\top U = \mathrm{Id}_m$, we have

$$\mathbb{E}\left[\sum_{i\in S}(U^i)(U^i)^\top\right] = \frac{1}{3}\cdot \mathrm{Id}_m\,.$$

By the assumption on the row norms of U, $\|U^i(U^i)^\top\|_2 = \|U^i\|^2 \le \mu\frac{m}{n}$. By the incoherence assumption, we have that $\max_i\|U^i\|^2 \le \mu m/n$.

We also note that U^i's are negatively associated random variables. Therefore by the matrix Chernoff inequality for negatively associated random variables, we have with high probability,

$$\left\|(U^S)^\top U^S - \mathbb{E}\left[(U^S)^\top U^S\right]\right\| \lesssim \sqrt{\frac{\mu m\log n}{n}}\,.$$

But, an analogous argument holds for S_b, S_c as well – so by a union bound over k, we complete the proof. □

Lemma 100. *Suppose $n \gtrsim m\log n$. Under the generative assumption in Section* **??** *for W, we have that we have that $F = 1 - \exp(-W)$ is $O(1)$-incoherent.*

Proof. We have that $FF^\top = U\Sigma^2U^T$ and therefore, $\|F^i\|^2 = \Sigma_{i,i}^2\|U^i\|^2$. This in turn implies that

$$\|U^i\|^2 \le \frac{1}{\min\Sigma_{ii}^2}\left\|F^i\right\| = \frac{1}{\sigma_{\min}^2(F)}\|F^i\|^2\,.$$

Since $\left\|F^i\right\|^2 \le pm + \sqrt{pm} \le 2pm$ with high probability, we only need to bound $\sigma_{\min}(F)$ from below. Note that $\sigma_{\min}^2(F) = \sigma_{\min}(F^\top F)$. Therefore it suffices to control $\sigma_{\min}(F^\top F)$. But by Lemma 98 we have $\sigma_{\min}^2(F) \gtrsim np$.

Therefore, we have that

$$\|U^i\|^2 \le \frac{1}{\sigma_{\min}^2(F)}\|F^i\|^2 = O(\frac{m}{n})$$

□

3.3.8 Spectral boundedness

The main goal of the section is to show that the bias terms in the PMI tensor are spectrally bounded by the PMI matrix (which we can estimate from samples). Furthermore, we show that we can calculate an approximate whitening matrix for the leading terms of the PMI tensor using the machinery in Section .

The main proposition we will show is the following:

Proposition 101. *Let W be sampled according to the random model in Section ?? with $\rho = o\left(\frac{1}{\log n}\right)$, $p = \omega\left(\frac{\log n}{\sqrt{mn}}\right)$. Let $S_a \subseteq [n]$, $|S_a| = \Omega(n)$. If R_{S_a} is the matrix that has as columns the vectors $\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)^{1/3}(\tilde{P}_l)_{j,S_a}$, $l \in [2, L]$, $j \in [m]$ for $L = O(poly(n))$, and A is the matrix that has as columns the vectors $\left(\frac{\rho}{1-\rho}\right)^{1/3}(\tilde{P}_1)_{j,S_a}$ for $j \in [m]$, with high probability it holds that $R_{S_a}R_{S_a}^\top$ is $O(\rho^{2/3}\log n)$-spectrally bounded by A.*

The main element for the proposition is the following Lemma:

Lemma 102. *For any set $S_a \subseteq [n]$, $|S_a| = \Omega(n)$, with high probability over the choice of W, for every $\ell, \ell' = O(poly(n))$ $P_{l,S_a}P_{l,S_a}^\top$ is $O(\log n)$-spectrally bounded by P_{l',S_a}.*

Before proving the Lemma, let us see how the proposition follows from it:

Proof of 101. We have

$$R_{S_a}R_{S_a}^\top = \sum_{l=2}^{L}\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)^{2/3}P_{l,S_a}$$

By Lemma 102, we have that $\forall l > 1$, $\tilde{P}_{l,S_a}$ is τ-spectrally bounded by $\tilde{P}_{1,S_a}$, for some $\tau = O(\log n)$. Hence,

$$R_{S_a}R_{S_a}^\top \preceq \sum_{l=2}^{\infty}\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)^{2/3}\tau(P_{1,S_a} + \sigma_{\min}(P_{1,S_a}))$$

$$\precsim \rho^{4/3}\tau(P_{1,S_a} + \sigma_{\min}(P_{1,S_a}))$$

Since $AA^\top = (\frac{\rho}{1-\rho})^{2/3}P_{1,S_a}$, the claim of the Proposition follows.

It is clear analogous statements hold for S_b and S_c. □

Finally, we proceed to show Lemma 102.

For notational convenience, we will denote by $J_{m\times n}$ the all ones matrix with dimension $m \times n$. (We will omit the dimensions when clear from the context.)

This statement will immediately follow from the following two lemmas:

Lemma 103. *For any set $S_a \subseteq [n]$, $|S_a| = \Omega(n)$, with probability $1 - \exp(-\log^2 n)$ over the choice of W, for all $\ell \leq O(poly(n))$,*

$$P_{l,S_a} \preceq 10np\log n\mathrm{Id} + \frac{5}{2}mp^2J$$

Lemma 104. *For any set $S_a \subseteq [n]$, $|S_a| = \Omega(n)$, with probability $1 - \exp(-\log^2 n)$ over the choice of W, $\forall \ell = poly(n)$,*

$$P_{l,S_a} + 6np \log n \mathrm{Id} \preccurlyeq mp^2 J$$

Before showing these lemmas, let us see how Lemma 102 is implied by them:

Proof of Lemma 102. Let κ be the constant in 104, s.t. $P_{l,S_a} + 6np \log n \mathrm{Id} \preccurlyeq mp^2 J$. Putting the bounds from Lemmas 103 and 104 together along with a union bound, we have that with high probability, $\forall l, l' = O(\mathrm{poly}(n))$

$$P_{l,S_a} - \frac{5}{2}\kappa P_{l',S_a} \preceq \left(10np \log n + \frac{15}{2}\kappa np \log n\right) \mathrm{Id} \preceq O(mp \log n)\mathrm{Id}$$

But, note that $\sigma_{\min}(P_{l'}) = \Omega(np)$, by Lemma 98. Hence, $P_l - \frac{5}{2}\kappa P_{l',S_a} \preceq r \log n \sigma_{\min}(P_{l'})$, for some sufficiently large constant r. This implies

$$P_l - r \log n P_{l',S_a} \preceq P_l - \frac{5}{2}\kappa P_{l'} \preceq r \log n \sigma_{\min}(P_{l',S_a})$$

from which the statement of the lemma follows.

□

We proceed to the first lemma:

Proof of Lemma 103. To make the notation less cluttered, we will drop l and S_a and use $P = P_{l,S_a}$. Furthermore, we will drop S_a when referring to columns of $\tilde{P}$ so we will denote $\tilde{P}_k = \tilde{P}_{k,S_a}$.

Let's denote by $e = \frac{1}{\sqrt{|S_a|}}\mathbf{1}$. Let's furthermore denote $\mathrm{Id}_{\mathbf{1}} = ee^\top$, and $\mathrm{Id}_{-\mathbf{1}} = \mathrm{Id} - ee^\top$. Note first that trivially, since $\mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{-\mathbf{1}} = \mathrm{Id}$,

$$P = (\mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{-\mathbf{1}})P(\mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{-\mathbf{1}}) \tag{3.3.33}$$

Furthermore, it also holds that

$$\begin{aligned} 0 &\preceq (2\mathrm{Id}_{-\mathbf{1}} - \frac{1}{2}\mathrm{Id}_{\mathbf{1}})P(2\mathrm{Id}_{-\mathbf{1}} - \frac{1}{2}\mathrm{Id}_{\mathbf{1}}) \\ &= \frac{1}{4}\mathrm{Id}_{\mathbf{1}}P\mathrm{Id}_{\mathbf{1}} + 4\mathrm{Id}_{-\mathbf{1}}P\mathrm{Id}_{-\mathbf{1}} - \mathrm{Id}_{-\mathbf{1}}P\mathrm{Id}_{\mathbf{1}} - \mathrm{Id}_{\mathbf{1}}P\mathrm{Id}_{-\mathbf{1}} \end{aligned}$$

where the first inequality holds since

$$(2\mathrm{Id}_{-\mathbf{1}} - \frac{1}{2}\mathrm{Id}_{\mathbf{1}})P(2\mathrm{Id}_{-\mathbf{1}} - \frac{1}{2}\mathrm{Id}_{\mathbf{1}}) = \left((2\mathrm{Id}_{-\mathbf{1}} - \frac{1}{2}\mathrm{Id}_{\mathbf{1}})\tilde{P}\right)\left((2\mathrm{Id}_{-\mathbf{1}} - \frac{1}{2}\mathrm{Id}_{\mathbf{1}})\tilde{P}\right)^\top$$

From this we get that

$$\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \leq \frac{1}{4} \mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} + 4 \mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \tag{3.3.34}$$

We proceed to upper bound both the terms on the RHS above. More precisely, we will show that

$$\mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} \leq 2mp^2 nJ \tag{3.3.35}$$

$$\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \leq 2np \log n \mathrm{Id} \tag{3.3.36}$$

Let us proceed to showing (3.3.35). The LHS can be rewritten as

$$\mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} = ee^\top \left(e^\top \tilde{P} \tilde{P}^\top e \right)$$

Note that

$$e^\top \tilde{P} \tilde{P}^\top e = \frac{1}{n} \left(\sum_{k=1}^{m} \langle \mathbf{1}, \tilde{P}_k \rangle^2 \right)$$

All the terms $\langle \mathbf{1}, \tilde{P}_k \rangle^2$ are independent and satisfy and $\mathbb{E}[\langle \mathbf{1}, \tilde{P}_k \rangle^2] \leq (\mathbb{E}[\langle \mathbf{1}, \tilde{P}_k \rangle])^2 \leq p^2 n^2$. By Chernoff, we have that

$$\sum_{k=1}^{m} \langle \mathbf{1}, \tilde{P}_k \rangle^2 \leq mp^2 n^2 + \sqrt{mp^2 n^2} \log n \leq 2mp^2 n^2$$

with probability at least $1 - \exp(-\log^2 n)$, where the second inequality holds because $p = \omega(\frac{\log n}{\sqrt{mn}})$. Hence, $e^\top \tilde{P} \tilde{P}^\top e \leq 2mp^2 n$ with high probability.

We proceed to (3.3.36), which will be shown by a Bernstein bound. Towards that, note that

$$\begin{aligned}
\mathbb{E}[(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k (\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)^\top] &= \mathrm{Id}_{-\mathbf{1}} \mathbb{E}[\tilde{P}_k (\tilde{P}_k)^\top] \mathrm{Id}_{-\mathbf{1}} \\
&= \mathrm{Id}_{-\mathbf{1}} \left(p^2 \mathbb{E}[1 - \exp(\tilde{W})]^2 \mathbf{1}\mathbf{1}^\top + (p - p^2) \mathbb{E}[(1 - \exp(\tilde{W}))^2] \mathrm{Id} \right) \mathrm{Id}_{-\mathbf{1}} \\
&\leq p \mathrm{Id}
\end{aligned}$$

where $\tilde{W}$ is a random variable following the distribution $\mathcal{D}$ of all the $\tilde{W}_{i,j}$. The second line can be seen to follow from the independence of the coordinates of $\tilde{P}_k$ according to our model.

Furthermore, with high probability $\|\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k\|_2^2 \leq \|\tilde{P}_k\|_2^2 \leq np$ and the random variable $\|\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k\|^2$ is sub-exponential

Finally,

$$r^2 = \left\| \mathbb{E}\left[\sum_{k=1}^{m} \left((\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)(\mathrm{Id}_{-\mathbf{1}} \Gamma_k)^{\tilde{P}} \right)^2 \right] \right\| \leq \left\| \tilde{P}_k \right\|_2^2 \left\| \mathbb{E}\left[\sum_{k=1}^{m} (I_{-\mathbf{1}} \tilde{P}_k)(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)^\top \right] \right\|$$
$$\leq npm \| \mathbb{E}[(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)^\top] \| \leq np^2 m$$

Therefore, applying a matrix Bernstein bound, we get

$$\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \leq mp\mathrm{Id} + np \log n \mathrm{Id} + r \log n \mathrm{Id} \leq 2np \log n \mathrm{Id}$$

with high probability.

Combining this with (3.3.33) and (3.3.34), we get

$$P \leq \frac{5}{4} \mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} + 5 \mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}}$$
$$\leq \frac{5}{2} mp^2 J + 10np \log n \mathrm{Id}$$

□

Let us proceed to the second inequality, which essentially follows the same strategy:

Proof of Lemma 104. Similarly as in the proof of Lemma 103, for notational convenience, let's denote by $\tilde{P}$ the matrix which has column k the vector $(1 - \exp(lW_k))$.

Reusing the notation from Lemma 104, we have that

$$P = (\mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{-\mathbf{1}}) P (\mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{-\mathbf{1}}) \tag{3.3.37}$$

and

$$0 \leq (\frac{1}{2} \mathrm{Id}_{\mathbf{1}} + 2\mathrm{Id}_{-\mathbf{1}}) P (\frac{1}{2} \mathrm{Id}_{\mathbf{1}} + 2\mathrm{Id}_{-\mathbf{1}})$$
$$= \mathrm{Id}_{-\mathbf{1}} P' \mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} + \frac{1}{4} \mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} + 4 \mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}}$$

for similar reasons as before. From this we get that

$$\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} + \mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \geq -\frac{1}{4} \mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} - 4 \mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \tag{3.3.38}$$

Putting (3.3.37) and (3.3.38) together, we get that

$$P + 3\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \succeq \frac{3}{4}\mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} \tag{3.3.39}$$

We will proceed to show an upper bound $\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \preceq 2np \log n \mathrm{Id}$ on second term of the LHS. We will do this by a Bernstein bound as before. Namely, analogously as in Lemma 104,

$$\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} = \sum_{k=1}^{m} (\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)^\top$$

and $\mathbb{E}[(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)^\top] \preceq p\mathrm{Id}$ and and $\|\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k\|_2^2 \leq \|\tilde{P}_k\|_2^2 \leq np$ are satisfied so

$$r^2 = \left\| \mathbb{E}\left[\sum_{k=1}^{m} \left((\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)(\mathrm{Id}_{-\mathbf{1}} \tilde{P}_k)^\top \right)^2 \right] \right\| \leq np^2 m$$

Therefore, applying a matrix Bernstein bound, we get

$$\mathrm{Id}_{-\mathbf{1}} P \mathrm{Id}_{-\mathbf{1}} \preceq mp\mathrm{Id} + np \log n \mathrm{Id} + r \preceq 2np \log n \mathrm{Id}$$

with high probability.

Plugging this in in (3.3.39), we get

$$P + 6np \log n \mathrm{Id} \succeq \frac{3}{4}\mathrm{Id}_{\mathbf{1}} P \mathrm{Id}_{\mathbf{1}} = \frac{3}{4} ee^\top \tilde{P} \tilde{P}^\top ee^\top = \frac{3}{4} ee^\top \left(e^\top \tilde{P} \tilde{P}^\top e \right)$$

Since we have $e^\top \tilde{P} \tilde{P}^\top e = \frac{1}{n}\left(\sum_{k=1}^{m} \langle \tilde{P}_k, \mathbf{1} \rangle^2 \right)$ with the goal of applying Chernoff, we will lower bound $\mathbb{E}[\langle \mathbf{1}, A_k \rangle^2]$. More precisely, we will show $\mathbb{E}[\langle \mathbf{1}, \tilde{P}_k \rangle]^2 = \Omega(n^2 p^2)$. In order to do this, we have

$$\begin{aligned}
\mathbb{E}[\langle \mathbf{1}, \tilde{P}_k \rangle^2] &= \sum_{j \in S_a} \mathbb{E}[(\tilde{P}_k)_j^2] + \sum_{j \neq j', j, j' \in S_a} \mathbb{E}[(\tilde{P}_k)_j] \mathbb{E}[(\tilde{P}_k)_{j'}] \\
&= \sum_{j \in S_a} p\mathbb{E}[(1 - \exp(-l\tilde{W})^2] + \sum_{j \neq j', j, j' \in S_a} p^2 \mathbb{E}[1 - \exp(-l\tilde{W})]^2 \\
&\geq \sum_{j \in S_a} p\mathbb{E}[(1 - \exp(-\tilde{W})^2] + \sum_{j \neq j', j, j' \in S_a} p^2 \mathbb{E}[1 - \exp(-\tilde{W})]^2 \\
&= \Omega(n^2 p^2)
\end{aligned}$$

where $\tilde{W}$ is a random variable following the distribution $\mathcal{D}$ of all the $\tilde{W}_{i,j}$. and the last inequality holds because of (3.3.2).

So by Chernoff, we get that $e^\top \tilde{P} \tilde{P}^\top e = \frac{1}{n}(\Omega(mn^2 p^2) - \sqrt{n^2 p^2 m}) = \Omega(mnp^2)$ with high probability. Altogether, this

means

$$P + 6np \log n \mathrm{Id} \gtrsim mp^2 J$$

□

3.3.9 Robust whitening

Algorithm 13 Obtaining whitening matrices

Inputs: Random partitioning S_a, S_b, S_c of $[n]$. Empirical PMI matrix $\widehat{\mathrm{PMI}}$.
Outputs: Whitening matrices $Q_a, Q_b, Q_c \in \mathbb{R}^{d \times d}$

1. Output

$$Q_a = \rho^{-1/3} \widehat{\mathrm{PMI}}_{S_a,S_b} (\widehat{\mathrm{PMI}}^{+}_{S_b,S_c})^\top \widehat{\mathrm{PMI}}_{S_c,S_a},$$
$$Q_b = \rho^{-1/3} \widehat{\mathrm{PMI}}_{S_b,S_c} (\widehat{\mathrm{PMI}}^{+}_{S_c,S_a})^\top \widehat{\mathrm{PMI}}_{S_a,S_b},$$
$$Q_c = \rho^{-1/3} \widehat{\mathrm{PMI}}_{S_c,S_a} (\widehat{\mathrm{PMI}}^{+}_{S_a,S_b})^\top \widehat{\mathrm{PMI}}_{S_b,S_c}$$

In this section, we show the formula $Q_a = \rho^{-1/3} \widehat{\mathrm{PMI}}_{S_a,S_b} (\widehat{\mathrm{PMI}}^{+}_{S_b,S_c})^\top \widehat{\mathrm{PMI}}_{S_c,S_a}$ computes an approximation of the true whitening matrix $AA^\top$, so that the error is ϵ-spectrally bounded by A. We recall Theorem 6.

@title. *Let $n \geq m$ and $A, B, C \in \mathbb{R}^{n \times m}$. Suppose $\Sigma_{ab}, \Sigma_{bc}, \Sigma_{ca} \in \mathbb{R}^{n \times n}$ are of the form,*

$$\Sigma_{ab} = AB^\top + E_{ab}, \quad \Sigma_{bc} = BC^\top + E_{bc}, \quad \text{and} \quad \Sigma_{ca} = CA^\top + E_{ca}.$$

where E_{ab}, E_{bc}, E_{ca} are ϵ-spectrally bounded by (A, B), (B, C), (C, A) respectively. Then, the matrix matrix

$$Q_a = \Sigma_{ab} [\Sigma_{bc}^\top]_m^+ \Sigma_{ca}$$

*is a good approximation of $AA^\top$ in the sense that $Q_a = \Sigma_{ab} [\Sigma_{bc}^\top]_m^+ \Sigma_{ca} - AA^\top$ is $O(\epsilon)$-*spectrally bounded *by A. Here $[\Sigma]_m$ denotes the best rank-m approximation of Σ.*

Towards proving Theorem 6, an intermediate step is to understand the how the space of singular vectors of $BC^\top$ are aligned with the noisy version Σ_{bc}. The following explicitly represent $BC^\top + E$ as the form $B'R(C')^\top + \Delta'$. Here the crucial benefit to do so is that the resulting Δ' is small in every direction. In other words, we started with a relative error guarantees on E and the Lemma below converts to it an absolute error guarantees on Δ' (though the signal term changes slightly).

Lemma 105. *Suppose B, C are $n \times m$ matrices with $n \geq m$. Suppose a matrix E is ϵ-spectrally bounded by (B, C),*

then $BC^\top + E$ *can be written as*

$$BC^\top + E = (B + \Delta_B)R_{BC}(C + \Delta_C)^\top + \Delta'_{BC}\,,$$

where $\Delta_B, \Delta_C, \Delta'_{BC}$ *are small and* R_{BC} *is close to identity in the sense that,*

$$\|\Delta_B\| \le O(\epsilon\sigma_{min}(B))$$
$$\|\Delta_C\| \le O(\epsilon\sigma_{min}(C))$$
$$\|\Delta'_{BC}\| \le O(\epsilon\sigma_{min}(B)\sigma_{min}(C))$$
$$\|R_{BC} - \mathrm{Id}\| \le O(\epsilon)$$

Proof. The key intuition is if the perturbation is happening in the span of columns of B and C, they cannot change the subspace. By Definition , we can write E as

$$E = B\Delta_1 C^\top + B\Delta_2^\top + \Delta_3 C^\top + \Delta_4.$$

Now since $\|\Delta_1\| \le \epsilon < 1$, we know $(\mathrm{Id} - \Delta_1)$ is invertible, so we can write

$$(BC^\top + E) = (B + \Delta_3(\mathrm{Id} + \Delta_1)^{-1})(\mathrm{Id} + \Delta_1)(C + \Delta_2(\mathrm{Id} - \Delta_1)^{-\top})^\top - \Delta_3(\mathrm{Id} - \Delta_1)^{-1}\Delta_2^\top + \Delta_4.$$

This is already in the desired form as we can let $\Delta_B = \Delta_3(\mathrm{Id} + \Delta_1)^{-1}$, $R_{BC} = (\mathrm{Id} + \Delta_1)$, $\Delta_C = \Delta_2(\mathrm{Id} - \Delta_1)^{-\top}$, and $\Delta'_{BC} = -\Delta_3(\mathrm{Id}-\Delta_1)^{-1}\Delta_2^\top + \Delta_4$. By Weyl's Theorem we know $\sigma_{min}(\mathrm{Id}+\Delta_1) \ge 1-\epsilon$, therefore $\|\Delta_B\| \le \|\Delta_3\|\sigma_{min}^{-1}(\mathrm{Id}+\Delta) \le \frac{\epsilon}{1-\epsilon}\sigma_{min}(B)$. Other terms can be bounded similarly.

Now we prove that the top m approximation of $BC^\top + E$ has similar column/row spaces as $BC^\top$. Let U_B be the column span of B, U'_B be the column span of $(B + \Delta_B)$, and U''_B be the top m left singular subspace of $(BC^\top + E)$. Similarly we can define U_C, U'_C, U''_C to be the column spans of C, $C + \Delta_C$ and the top m right singular subspace of $(BC^\top + E)$.

For $B + \Delta_B$, we can apply Weyl's Theorem and Wedin's Theorem. By Weyl's Theorem we know $\sigma_{min}(B + \Delta_B) \ge \sigma_{min}(B) - \|\Delta_B\| \ge (1 - O(\epsilon))\sigma_{min}(B)$. By Wedin's Theorem we know U'_B is $O(\epsilon)$-close to U_B. Similar results apply to $C + \Delta_C$.

Now we know $\sigma_{min}((B+\Delta_B)R_{BC}(C+\Delta_C)^\top)) \ge \sigma_{min}(B+\Delta_B)\sigma_{min}(R_{BC})\sigma_{min}(C+\Delta_C) \ge \Omega(\sigma_{min}(B)\sigma_{min}(C)$. Therefore we can again apply Wedin's Theorem, considering $(B + \Delta_B)R_{BC}(C + \Delta_C)^\top)$ as the original matrix and Δ'_{BC} as the perturbation. As a result, we know U''_B is $O(\epsilon)$ close to U'_B, U''_C is $O(\epsilon)$ close to U'_B. The distance between U_B, U''_B (and U_C, U''_C) then follows from triangle inequality.

□

As a direct corollary of Lemma 105, we obtain that the $BC^\top$ and $BC^\top + E$ have similar subspaces of singular vectors.

Corollary 106. *In the setting of Lemma 105, let $[BC^\top + E]_m$ be the best rank-m approximation of $BC^\top + E$. Then, the span of columns of $[BC^\top + E]_m$ is $O(\epsilon)$-close to the span of columns of B, span of rows of $[BC^\top + E]_m$ is $O(\epsilon)$-close to the span of columns of C.*

Furthermore, we can write $[BC^\top + E]_m = (B + \Delta_B)R_{BC}(C + \Delta_C)^\top + \Delta_{BC}$. Here Δ_B, Δ_C and R_{BC} as defined in Lemma 105 and Δ_{BC} satisfies $\|\Delta_{BC}\| \le O(\epsilon\sigma_{min}(B)\sigma_{min}(C))$.

Proof. Since $[BC^\top + E]_m$ is the best rank-m approximation, because $(B + \Delta_B)R_{BC}(C + \Delta_C)^\top)$ is a rank m matrix, in particular we have

$$\|BC^\top + E - [BC^\top + E]_m\| \le \|BC^\top - (B + \Delta_B)R_{BC}(C + \Delta_C)^\top)\|\|\Delta'_{BC}\|.$$

Therefore

$$\begin{aligned}\|\Delta_{BC}\| &= \|[BC^\top + E]_m - (B + \Delta_B)R_{BC}(C + \Delta_C)^\top)\| \\ &\le \|BC^\top + E - [BC^\top + E]_m\| + \|BC^\top + E - (B + \Delta_B)R_{BC}(C + \Delta_C)^\top)\| \\ &\le 2\|\Delta'_{BC}\|.\end{aligned}$$

□

In order to fix this problem, we notice that the matrix $[\Sigma_{bc}^\top]_m^+$ is multiplied by Σ_{ab} on the left and Σ_{ca} on the right. Assuming $\Sigma_{ab} = AB^\top$, $\Sigma_{ca} = CA^\top$, we should expect $[\Sigma_{bc}^\top]_m^+$ to "cancel" with the $B^\top$ factor on the left and the C factor on the right, giving us $AA^\top$. Therefore, we should really measure the error of the middle term $[\Sigma_{bc}^\top]_m^+$ after left multiplying with $B^\top$ and right multiplying with C. We formalize this in the following lemma:

Lemma 107. *Suppose Σ_{bc} is as defined in Theorem 6, let $\Delta = [\Sigma_{bc}^\top]_m^+ - [CB^\top]^+$, then we have*

$$\|B^\top\Delta C\| = O(\epsilon), \quad \|B^\top\Delta\| \le O(\frac{\epsilon}{\sigma_{min}(C)}),$$
$$\|\Delta C\| \le O(\frac{\epsilon}{\sigma_{min}(B)}), \quad \|\Delta\| \le O(\frac{\epsilon}{\sigma_{min}(B)\sigma_{min}(C)}).$$

We will first prove Theorem 6 assuming Lemma 107.

Proof of Theorem 6. By Lemma 105, we know Σ_{ab} can be written as

$$(A + \Delta_A^1)R_{AB}(B + \Delta_B^1)^\top + \Delta_{AB}.$$

Similarly Σ_{ca} can be written as

$$(C + \Delta_C^3)R_{CA}(A + \Delta_A^3)^\top + \Delta_{CA}.$$

Here the Δ terms and R terms are bounded as in Lemma 105.

Now let us write the matrix $\Sigma_{ab}[\Sigma_{bc}^\top]_m^+\Sigma_{ca}$ as

$$\left((A + \Delta_A^1)R_{AB}(B + \Delta_B^1)^\top + \Delta_{AB}\right)([CB^\top]^+ + \Delta_{BC})\left((C + \Delta_C^3)R_{CA}(A + \Delta_A^3)^\top + \Delta_{CA}\right)$$

We can now view $\Sigma_{ab}[\Sigma_{bc}^\top]_m^+\Sigma_{ca}$ as the product of three terms, each term is the sum of two matrices. Therefore we can expand the product into 8 terms. In each of the three pairs, we will call the first matrix the main matrix, and the second matrix the perturbation.

In the remaining proof, we will do calculations to show the product of the main terms is close to $AA^\top$, and all the other 7 terms are small.

Before doing that, we first prove several Claims about PSD matrices

Claim 11. *If $\|\Delta\| \le \epsilon$, then $A\Delta A^\top \preceq \epsilon AA^\top$. If $\|\Gamma\| \le \epsilon\sigma_{min}(A)$, then $\frac{1}{2}(A\Gamma^\top + \Gamma A^\top) \preceq \epsilon AA^\top + \epsilon\sigma_{min}^2(A)\mathrm{Id}$.*

Proof. Both inequalities can be proved by consider the quadratic form. We know for any x, $x^\top A\Delta A^\top x \le \|\Delta\|\|A^\top x\|^2 \le \epsilon x^\top AA^\top x$, so the first part is true.

For the second part, for any x we can apply Cauchy-Schwartz inequality

$$x^\top\frac{1}{2}(A\Gamma^\top + \Gamma A^\top)x = \langle\sqrt{\epsilon}A^\top x, \epsilon^{-1/2}\Gamma^\top x\rangle \le \epsilon\|A^\top x\|^2 + \epsilon^{-1}\|\Gamma^\top x\|^2 = x^\top(\epsilon AA^\top + \epsilon\sigma_{min}^2(A)\mathrm{Id})x.$$

□

Now, we will first prove the product of three main matrices is close to $AA^\top$:

Claim 12. *We have $\left((A + \Delta_A^1)R_{AB}(B + \Delta_B^1)^\top\right)(CB^\top)^+\left((C + \Delta_C^3)R_{CA}(A + \Delta_A^3)^\top\right) = AA^\top + E_A$, where E_A is $O(\epsilon)$-spectrally bounded by $AA^\top$.*

Proof. We will first prove the middle part of the matrix $(B + \Delta_B^1)^\top(CB^\top)^+(C + \Delta_C^3)$ is $O(\epsilon)$ close to identity matrix Id. Here we observe that both B, C have full column rank so $(CB^\top)^+ = (B^\top)^+C^+$. Therefore we can rewrite the product as $(\mathrm{Id} + B^+\Delta_B^1)^\top(\mathrm{Id} + C^+\Delta_C^3)$. Since $\|\Delta_B^1\| \le O(\epsilon\sigma_{min}(B))$ by Lemma 105 (and similarly for C), we know $\|B^+\Delta_B^1\| \le O(\epsilon)$. Therefore the middle part is $O(\epsilon)$ close to Id. Now since $\epsilon \ll 1$ we know $\widehat{R}_{AB} = R_{AB}(B + \Delta_B^1)^\top(CB^\top)^+(C + \Delta_C^3)R_{CA}$ is $O(\epsilon)$-close to Id.

Now we are left with $(A+\Delta_A^1)\widehat{R}_{AB}(A+\Delta_A^3)^\top$, for this matrix we know

$$(A+\Delta_A^1)\widehat{R}_{AB}(A+\Delta_A^3)^\top - AA^\top = A(\widehat{R}_{AB}-\mathrm{Id})A^\top + \Delta_A^1\widehat{R}_{AB}A^\top + A\widehat{R}_{AB}(\Delta_A^3)^\top + \Delta_A^1\widehat{R}_{AB}(\Delta_A^3)^\top.$$

The first term $A(\widehat{R}_{AB}-\mathrm{Id})A^\top \preceq O(\epsilon)AA^\top$ (Claim 11); the fourth term $\Delta_A^1\widehat{R}_{AB}(\Delta_A^3)^\top \preceq O(\epsilon\sigma_{min}^2(A))\mathrm{Id}$ (by the norm bounds of Δ_A^1 and Δ_A^3. For the cross terms, we can bound them using the second part of Claim 11. □

Next we will try to prove the remaining 7 terms are small. We partition them into three types depending on how many Δ factors they have. We proceed to bound them in each of these cases.

For the terms with only one Δ, we claim:

Claim 13. *The three terms* $\Delta_{AB}(CB^\top)^+(C+\Delta_C^3)R_{CA}(A+\Delta_A^3)^\top$, $(A+\Delta_A^1)R_{AB}(B+\Delta_B^1)^\top\Delta_{AB}(C+\Delta_C^3)R_{CA}(A+\Delta_A^3)^\top$, $\left((A+\Delta_A^1)R_{AB}(B+\Delta_B^1)^\top\right)(CB^\top)^+\Delta_{CA}$ *are all* $O(\epsilon)$ *spectrally bounded by* $AA^\top$.

Proof. For the first term, note that both B, C have full column rank, and hence $(CB^\top)^+ = (B^\top)^+C^+$. Therefore the first term can be rewritten as

$$[\Delta_{AB}(B^\top)^+][(\mathrm{Id}+C^+\Delta_C^3)R_{CA}](A+\Delta_A^3)^\top.$$

By Lemma 105, we have spectral norm bounds for $\Delta_{AB}, \Delta_C^3, \Delta_A^3, R_{CA}$. Therefore we know $\|\Delta_{AB}(B^\top)^+\| \le O(\epsilon\sigma_{min}(A))$ and $[(\mathrm{Id}+C^+\Delta_C^3)R_{CA}]$ is $O(\epsilon)$ close to Id. Therefore $\|[\Delta_{AB}(B^\top)^+][(\mathrm{Id}+C^+\Delta_C^3)R_{CA}](\Delta_A^3)^\top\| \le O(\epsilon\sigma_{min}^2(A))$ is trivially $O(\epsilon)$ spectrally bounded, and $[\Delta_{AB}(B^\top)^+][(\mathrm{Id}+C^+\Delta_C^3)R_{CA}]A^\top$ is $O(\epsilon)$ spectrally bounded by Claim 11. The third term is exactly symmetric.

For the second part, we will first prove the middle part of the matrix $\widehat{\Delta}_{BC} = (B+\Delta_B^1)^\top\Delta_{BC}(C+\Delta_C^3)$ has spectral norm $O(\epsilon)$. This can be done y expanding it to the sum of 4 terms, and use appropriate spectral norm bounds on Δ_{BC} and its products with $B^\top$ and C from Lemma 107. Now we can show $(A+\Delta_A^1)R_{AB}\widehat{\Delta}_{BC}R_{CA}(A+\Delta_A^3)^\top$ is $O(\epsilon)$ spectrally bounded by the first part of Claim 11. □

Next we try to bound the terms with two Δ factors.

Claim 14. *The three terms* $\Delta_{AB}\Delta_{BC}(C+\Delta_C^3)R_{CA}(A+\Delta_A^3)^\top$, $\Delta_{AB}(CB^\top)^+\Delta_{CA}$, $\left((A+\Delta_A^1)R_{AB}(B+\Delta_B^1)^\top\right)\Delta_{BC}\Delta_{CA}$ *are all* $O(\epsilon^2)$ *spectrally bounded by* $AA^\top$.

Proof. For the first term, notice that $\|\Delta_{BC}(C+\Delta_C^3)\|$ is bounded by $O(\epsilon/\sigma_{min}(B))$ by Lemma 107, and $\|\Delta_{AB}\| = O(\epsilon\sigma_{min}(A)\sigma_{min}(B))$. Therefore we know $\|\Delta_{AB}\Delta_{BC}(C+\Delta_C^3)R_{CA}\| \le O(\epsilon^2\sigma_{min}(A))$, so by Claim 11 we know this term is $O(\epsilon^2)$ spectrally bounded by $AA^\top$. Third term is symmetric.

For the second term, by Lemma 105 we can directly bound its spectral norm by $O(\epsilon^2\sigma_{min}^2(A))$, so it is trivially $O(\epsilon^2)$ spectrally bounded by $AA^\top$. □

Finally, for the product $\Delta_{AB}\Delta_{BC}\Delta_{CA}$, we can get the spectral norm for the three factors by Lemma 105 and Lemma 107. As a result $\|\Delta_{AB}\Delta_{BC}\Delta_{CA}\| \leq O(\epsilon^3\sigma^2_{min}(A))$ which is trivially $O(\epsilon^3)$ spectrally bounded by $AA^\top$.

Combining the bound for all of the terms we get the theorem. □

With that, we now try to prove Lemma 107

We first prove a simpler version where the perturbation is simply bounded in spectral norm

Lemma 108. *Suppose B, C are $n \times m$ matrices and $n \geq m$. Let R be an $n \times n$ matrix such that $\|R - \mathrm{Id}\| \leq \epsilon$, and E is a perturbation matrix with $\|E\| \leq \epsilon\sigma_{min}(B)\sigma_{min}(C)$ and $(CRB^\top + E)$ is also of rank m.*

Now let $\Delta = (CRB^\top + E)^+ - (CRB^\top)^+$, then when $\epsilon \ll 1$ we have

$$\|B^\top \Delta C\| = O(\epsilon), \quad \|B^\top \Delta\| \leq O(\frac{\epsilon}{\sigma_{min}(C)}),$$
$$\|\Delta C\| \leq O(\frac{\epsilon}{\sigma_{min}(B)}), \quad \|\Delta\| \leq O(\frac{\epsilon}{\sigma_{min}(B)\sigma_{min}(C)}).$$

Proof. We first give the proof for $\|B^\top \Delta C\|$. Other terms are similar.

Let U_B be the column span of B, and U'_B be the row span of $(CRB^\top + E)$. Similarly let U_C be the column span of C and U'_C be the column span of $(CRB^\top + E)$. By Wedin's theorem, we know U'_B is $O(\epsilon)$ close to U_B and U'_C is $O(\epsilon)$ close to U_C. As a result, suppose the SVD of B is $U_B D_B V_B^\top$, we know

$$\sigma_{min}(B^\top U'_B) = \sigma_{min}(V_B D_B U_B^\top U'_B) \geq (1 - O(\epsilon))\sigma_{min}(B).$$

The same is true for C: $\sigma_{min}(C^\top U'_C) \geq (1 - O(\epsilon))\sigma_{min}(C)$.

By the property of pseudoinverse, the column span of $(CRB^\top + E)^+$ is U'_B, and the row span of $(CRB^\top + E)^+$ is U'_C, further, $(CRB^\top + E)^+ = U'_B[(U'_C)^\top(CRB^\top + E)U'_B]^{-1}U'_C$, therefore we can write

$$B^\top(CRB^\top + E)^+C = B^\top U'_B[(U'_C)^\top(CRB^\top + E)U'_B]^{-1}(U'_C)^\top C.$$

Note that now the three matrices are all $n \times n$ and invertible! We can write $B^\top U'_B = ((B^\top U'_B)^{-1})^{-1}$ (and do the same thing for $(U'_C)^\top C$. Using the fact that $P^{-1}Q^{-1} = (QP)^{-1}$, we have

$$\begin{aligned} B^\top(CRB^\top + E)^+C &= (((U'_C)^\top C)^{-1}(U'_C)^\top(CRB^\top + E)U'_B(B^\top U'_B)^{-1})^{-1} \\ &= (R + ((U'_C)^\top C)^{-1}(U'_C)^\top E U'_B(B^\top U'_B)^{-1})^{-1} =: (R + X)^{-1}. \end{aligned}$$

Here we defined $X = ((U'_C)^\top C)^{-1}(U'_C)^\top E U'_B (B^\top U'_B)^{-1}$. The spectral norm of X can be bounded by

$$\begin{aligned}\|X\| &\le \|((U'_C)^\top C)^{-1}\|\|E\|\|(B^\top U'_B)^{-1}\| \\ &= \|E\|\sigma_{min}^{-1}(B^\top U'_B)\sigma_{min}^{-1}(C^\top C'_B) \\ &\le O(\epsilon).\end{aligned}$$

We can write $B^\top \Delta C = B^\top (CRB^\top + E)^+ C - \mathrm{Id} = (\mathrm{Id} + (R - \mathrm{Id} + X))^{-1} - \mathrm{Id}$, and we now know $\|(R - \mathrm{Id} + X)\| \le O(\epsilon)$, as a result $\|B^\top \Delta C\| \le O(\epsilon)$ as desired.

For the term $\|B^\top \Delta\|$, by the same argument we have

$$\begin{aligned}B^\top (CRB^\top + E)^+ &= ((U'_C)^\top (CRB^\top + E) U'_B (B^\top U'_B)^{-1})^{-1}(U'_C)^\top \\ &= ((U'_C)^\top CR + (U'_C)^\top E U'_B (B^\top U'_B)^{-1})^{-1}(U'_C)^\top \\ &= ((U'_C)^\top C(R + X))^{-1}(U'_C)^\top \\ &= (R + X)^{-1}((U'_C)^\top C)^{-1}(U'_C)^\top.\end{aligned}$$

On the other hand, we know $B^\top (CRB^\top)^+ = R^{-1}C^+ = R^{-1}((U_C)^\top C)^{-1}U_C^\top$. We can match the three factors:

$$\begin{aligned}\|R^{-1} - (R + X)^{-1}\| \le O(\epsilon), \quad \|R^{-1}\| \le 1 + O(\epsilon) \\ \|((U_C)^\top C)^{-1} - ((U'_C)^\top C)^{-1}\| \le O(\epsilon/\sigma_{min}(C)), \quad \|((U_C)^\top C)^{-1}\| = O(1/\sigma_{min}(C)) \\ \|U_C - U'_C\| \le O(\epsilon), \quad \|U_C\| = 1.\end{aligned}$$

Here, first and third bound are proven before. The second bound comes if we consider the SVD of $C = U_C D_C V_C^\top$ and notice that $\|(U'_C)^\top U_C - \mathrm{Id}\| \le O(\epsilon)$. We can write $\Delta_1 = R^{-1} - (R+X)^{-1}$, $\Delta_2 = ((U_C)^\top C)^{-1} - ((U'_C)^\top C)^{-1}$, $\Delta_3 = U_C - U'_C$, then we have

$$\begin{aligned}B^\top \Delta &= B^\top (CRB^\top + E)^+ - B^\top (CRB^\top)^+ \\ &= (R^{-1} - \Delta_1)(((U_C)^\top C)^{-1} - \Delta_2)(U_C - \Delta_3)^\top - R^{-1}((U_C)^\top C)^{-1}U_C^\top.\end{aligned}$$

Expanding the last equation, we get 7 terms and all of them can be bounded by $O(\epsilon/\sigma_{min}(C))$. The bounds on $\|\Delta C\|$ and $\|\Delta\|$ can be proved using similar techniques. □

Finally we are ready to prove the main Lemma 107:

Proof of Lemma 107. Using Lemma 105, let $E = E_{bc}^{\top}$, we can write the matrix before pseudoinverse as

$$[CB^{\top} + E]_m = (C + \Delta_C)R_{BC}(B + \Delta_B)^{\top} + \Delta_{BC}.$$

We can then apply Lemma 108 on $(C + \Delta_C)R_{BC}(B + \Delta_B)^{\top} + \Delta_{BC}$. As a result, we know if we let $\Delta' = [CB^{\top} + E]_m^+ - ((C + \Delta_C)R_{BC}(B + \Delta_B)^{\top})^+$, we have the desired bound if we left multiply with $(B + \Delta_B)^{\top}$ or right multiply with $(C + \Delta_C)$.

We will now show how to prove the first bound, all the other bounds can be proved using the same strategy:

First, we can write

$$B^{\top}\Delta' C = -(B + \Delta_B)^{\top}\Delta'(C + \Delta_C) + (B + \Delta_B)^{\top}\Delta' C + \Delta_B^{\top}\Delta'(C + \Delta_C) - \Delta_B^{\top}\Delta'\Delta_C.$$

All the four terms on the RHS can be bounded by Lemma 108 so we know $\|B^{\top}\Delta' C\| \leq O(\epsilon)$.

On the other hand, let $\Delta'' = ((C + \Delta_C)R_{BC}(B + \Delta_B)^{\top})^+ - (C^{\top}B)^+ = \Delta - \Delta'$. We will prove $\|B^{\top}\Delta'' C\| \leq O(\epsilon)$ and then the bound on $\|B^{\top}\Delta C\|$ follows from triangle inequality.

For $B^{\top}\Delta'' C$, we know it is equal to

$$B^{\top}[(B + \Delta_B)^{\top}]^+ R_{AB}^{-1}(C + \Delta_C)^+ C - \mathrm{Id}$$

Claim 15. $\|B^{\top}[(B + \Delta_B)^{\top}]^+ R_{AB}^{-1}(C + \Delta_C)^+ C - \mathrm{Id}\| \leq O(\epsilon)$

Proof. We will show all three factors in the first term are $O(\epsilon)$ close to Id. For R_{AB}^{-1} this follows immediately from Lemma 105. For $(C + \Delta_C)^+ C$, we know

$$(C + \Delta_C)^+ C - \mathrm{Id} = -(C + \Delta_C)^+ \Delta_C.$$

Therefore its spectral norm bound is bounded by $\|\Delta_C\|\sigma_{min}^{-1}(C + \Delta_C) = O(\epsilon)$ (where the bound on $\|\Delta_C\|$ comes from Lemma 105). □

With the claim we have now proven $\|B^{\top}\Delta'' C\| \leq O(\epsilon)$, therefore

$$\|B^{\top}\Delta C\| \leq \|B^{\top}(\Delta' + \Delta'')C\| \leq \|B^{\top}\Delta' C\| + \|B^{\top}\Delta'' C\| \leq O(\epsilon).$$

□

3.3.10 Technical details: spectral boundedness and incoherence

Here we will show under mild incoherence conditions (defined below), if an error matrix E is ϵ-spectrally bounded by $FF^\top$, then the partial matrices satisfy the requirement of Theorem 6.

Theorem 16. *If F is μ-incoherent for $\mu \leq \sqrt{n/m\log^2 n}$, then when $n \geq \Omega(m\log^2 m)$, with high probability over the random partition of F into A, B, C, we know $\sigma_{min}(A) \geq \sigma_{min}(F)/3$ (same is true for B, C).*

As a corollary, if E is ϵ-spectrally bounded by F. Let a, b, c be the subsets corresponding to A, B, C, and let $E_{a,b}$ be the submatrix of E whose rows are in set a and columns are in set b. Then $E_{a,b}$ (also $E_{b,c}, E_{c,a}$) is $O(\epsilon)$-spectrally bounded by the corresponding asymmetric matrices $AB^\top$ ($BC^\top$, $CA^\top$).

Proof. Consider the singular value decomposition of F: $F = UDV^\top$. Here U is a $n \times m$ matrix whose columns are orthonormal, V is an $m \times m$ orthonormal matrix and D is a diagonal matrix whose smallest diagonal entry is $\sigma_{min}(F)$.

Consider the following way of partitioning the matrix: for each row of F, we put it into A, B or C with probability $1/3$ independently.

Now, let $X_i = 1$ if row i is in the matrix A, and 0 otherwise. Then X_i's are Bernoulli random variables with probability $1/2$. Suppose S is the set of rows in A, let U_A be U restricted to rows in A, then we have $A = U_A DV^\top$. We will show with high probability $\sigma_{min}(A) \geq 1/3$.

The key observation here is the expectation of $U_A^\top U_A = \sum_{i=1}^n X_i U_i U_i^\top$, where U_i is the i-th row of U (represented as a column vector). Since X_i's are Bernoulli random variables, we know

$$\mathbb{E}[U_A^\top U_A] = \mathbb{E}[\sum_{i=1}^n X_i U_i U_i^\top] = \frac{1}{3}\sum_{i=1}^n U_i U_i^\top = \frac{1}{3}\mathrm{Id}.$$

Therefore we can hope to use matrix concentration to prove that $U_A^\top U_A$ is close to its expectation.

Let $M_i = X_i U_i U_i^\top - 1/3 U_i U_i^\top$. Clearly $\mathbb{E}[M_i] = 0$. By the Incoherence assumption, we know $\|U_i\| \leq 1/\log n$. Therefore we know $\|M_i\| \leq O(1/\log n)$. Also, we can bound the variance

$$\|\mathbb{E}[\sum_{i=1}^n M_i M_i^\top]\| \leq \|\mathbb{E}[\sum_{i=1}^n X_i U_i U_i^\top U_i U_i^\top]\| \leq \max\|U_i\|^2 \|\sum_{i=1}^n U_i U_i^\top\| \leq O(1/\log^2 n).$$

Here the last inequality is because $\sum_{i=1}^n X_i U_i U_i^\top U_i U_i^\top \preceq \|U_i\|^2 U_i U_i^\top$.

Therefore by Matrix Bernstein's inequality we know with high probability $\|\sum_{i=1}^n M_i\| \leq 1/6$. When this happens we know

$$\|U_A^\top U_A\| \geq \sigma_{min}(\mathbb{E}[U_A^\top U_A]) - \|\sum_{i=1}^n M_i\| \geq 1/6.$$

Hence we have $\sigma_{min}(U_A) \geq \sqrt{1/6} > 1/3$, and $\sigma_{min}(A) \geq \sigma_{min}(U_A)\sigma_{min}(D) \geq \sigma_{min}(F)/3$. Note that matrices B, C have exactly the same distribution as A so the bounds for B,C follows from union bound.

For the corollary, if a matrix E is ϵ spectrally bounded, we can write it as $F\Delta_1 F^\top + F\Delta_2^\top + \Delta_2 F^\top + \Delta_4$, where $\|\Delta_1\| \le \epsilon$, $\|\Delta_2\| \le \epsilon\sigma_{min}(F)$ and $\|\Delta_4\| \le \epsilon\sigma_{min}^2(F)$. This can be done by considering different projections of E: let U be the span of columns of F, then $F\Delta_1 F^\top$ term corresponds to $\mathrm{Proj}_U E \mathrm{Proj}_U$; $F\Delta_2^\top$ term corresponds to $\mathrm{Proj}_U E \mathrm{Proj}_{U^\perp}$; $\Delta_2 F^\top$ term corresponds to $\mathrm{Proj}_{U^\perp} E \mathrm{Proj}_U$; Δ_4 term corresponds to $\mathrm{Proj}_{U^\perp} E \mathrm{Proj}_{U^\perp}$. The spectral bounds are necessary for E to be spectrally bounded.

Now for $E_{a,b}$, we can write it as $A\Delta_1 B^\top + A(\Delta_2)_b^\top + (\Delta_2)_a B^\top + (\Delta_4)_{a,b}$, where we also take the corresponding submatrices of Δ's. Since the spectral norm of a submatrix can only be smaller, we know $\|\Delta_1\| \le \epsilon$, $\|(\Delta_2)_b\| \le \epsilon\sigma_{min}(F) \le 3\epsilon\sigma_{min}(B)$, $\|(\Delta_2)_a\| \le \epsilon\sigma_{min}(F) \le 3\epsilon\sigma_{min}(A)$ and $\|(\Delta_2)_{a,b}\| \le \epsilon\sigma_{min}^2(F) \le 9\epsilon\sigma_{min}(A)\sigma_{min}(B)$. Therefore by Definition we know $E_{a,b}$ is 9ϵ spectrally bounded by $AB^\top$. □

3.3.11 Putting things together: proof of Theorem 7 and Theorem 8

In this section, we provide the full proof of Theorem 7. We start with a simple technical Lemma.

Lemma 109. *If Q is an ϵ-approximate whitening matrix for A, then $\|Q\| \le \frac{1}{1-\epsilon}\|AA^\top\|$, $\sigma_{\min}(Q) \ge \frac{1}{1-\epsilon}\|AA^\top\|$*

Proof. By the definition of approximate-whitening, we have

$$1-\epsilon \le \sigma_{\min}((Q^+)^{1/2}A^\top A(Q^+)^{1/2}), \sigma_{\max}((Q^+)^{1/2}A^\top A(Q^+)^{1/2}) \le 1+\epsilon$$

which implies that

$$1-\epsilon \le \sigma_{\min}((Q^+)^{1/2}AA^\top(Q^+)^{1/2}), \sigma_{\max}((Q^+)^{1/2}AA^\top(Q^+)^{1/2}) \le 1+\epsilon$$

by virtue of the fact that $(Q^+)^{1/2}AA^\top(Q^+)^{1/2} = \left((Q^+)^{1/2}A^\top A(Q^+)^{1/2}\right)^\top$. Rewriting in semidefinite-order notation, we get that

$$(1-\epsilon)\mathrm{Id} \le (Q^+)^{1/2}AA^\top(Q^+)^{1/2} \le (1+\epsilon)\mathrm{Id}$$

Multiplying on the left and right by $Q^{1/2}$, we get

$$(1-\epsilon)Q \le AA^\top \le (1+\epsilon)Q$$

This directly implies $\frac{1}{1+\epsilon}AA^\top \le Q \le \frac{1}{1-\epsilon}AA^\top$ which is equivalent to the statement of the lemma. □

Towards proving Theorem 7, we will first prove the following proposition, which shows that we recover the $\exp(-W)$ matrix correctly:

Proposition 110 (Recovery of $\exp(-W)$)**.** *Under the random generative model defined in Section* **??***, if the number of samples satisfies*

$$N = \text{poly}(n, 1/p/, 1/\rho)$$

the vectors $\tilde{W}_i, i \in [m]$ *in Algorithm 17 are* $O(\eta\sqrt{np})$*-close to* $\exp(-W_i)$ *where*

$$\eta = \tilde{O}\left(\sqrt{mp}\rho\right)$$

Proof. The proof will consist of checking the conditions for Algorithms 13 and 12 to work, so that we can apply Theorems 10 and 6.

To get a handle on the PMI tensor, by Proposition 96, for any equipartition S_a, S_b, S_c of $[n]$, we can it as

$$\text{PMIT}_{S_a,S_b,S_c} =$$
$$\frac{\rho}{1-\rho}\sum_{k\in[m]} F_{k,S_a}\otimes F_{k,S_b}\otimes F_{k,S_c} + \sum_{l=2}^{L}(-1)^{l+1}\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)\sum_{k\in[m]}(\tilde{P}_l)_{k,S_a}\otimes(\tilde{P}_l)_{k,S_b}\otimes(\tilde{P}_l)_{k,S_c} + E_L$$

We can choose $L = \text{poly}(\log(n, \frac{1}{\rho}, \frac{1}{p}))$ to ensure

$$\|E_L\| = o\left(p^{5/2}\rho^{7/3}\sqrt{m}n^2\right) \tag{3.3.40}$$

Having an explicit form for the tensor, we proceed to check the spectral boundedness condition for Algorithm 13.

Let S_a, S_b, S_c be a random equipartition. Let R_{S_a} be the matrix that has as columns the vectors $\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)^{1/3}(\tilde{P}_l)_{j,S_a}$, for all $l \in [2, L]$, $j \in [m]$ and let A be the matrix that has as columns the vectors $\left(\frac{\rho}{1-\rho}\right)^{1/3} F_{j,S_a}$, for all $j \in [m]$. Since L is polynomially bounded in n, by Proposition 101 we have that with high probability R_{S_a} is τ-spectrally bounded by A, for a $\tau = O(\rho^{2/3}\log n)$. Analogous statements hold for S_b, S_c.

Next, we verify the conditions for calculating approximate whitening matrices (Algorithm 13).

Towards applying Theorem 16, note that if $R_{[n]}$ is the matrix that has as columns the vectors $\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)^{1/3}(\tilde{P}_l)_j$, for all $l \in [2, L]$, $j \in [m]$, and D is the matrix that has as columns the vectors $\left(\frac{\rho}{1-\rho}\right)^{1/3} F_j$, for all $j \in [m]$, $R_{[n]}$ then is τ-spectrally bounded by D for $\tau = O(\rho^{2/3}\log n)$. Furthermore, the matrix F is $O(1)$-incoherent with high probability by Lemma 3.3.7. Hence, we can apply Theorem 16, the output of Algorithm 13 are matrices Q_a, Q_b, Q_c which are τ-approximate whitening matrices for A, B, C respectively.

Next, we will need bounds on

$$\min(\sigma_{\min}(Q_a), \sigma_{\min}(Q_b), \sigma_{\min}(Q_c)), \max(\sigma_{\max}(Q_a), \sigma_{\max}(Q_b), \sigma_{\max}(Q_c))$$

to plug in the guarantee of Algorithm 13.

By Lemma 109, we have

$$\sigma_{\max}(Q_a) \leq \frac{1}{1-\tau}\|AA^\top\| \lesssim (1+\tau)\|AA^\top\|, \sigma_{\min}(Q_a) \geq \frac{1}{1+\tau}\sigma_{\min}(AA^\top) \gtrsim (1-\tau)\sigma_{\min}(AA^\top)$$

However, for the random model, applying Lemma 98,

$$\sigma_{\min}(AA^\top) \geq \left(\frac{\rho}{1-\rho}\right)^{2/3} np \gtrsim \rho^{2/3} np, \sigma_{\max}(AA^\top) \leq \left(\frac{\rho}{1-\rho}\right)^{2/3} mnp^2 \lesssim \rho^{2/3} mnp^2 \tag{3.3.41}$$

Analogous statements hold for B and C.

Finally, we bound the error due to empirical estimates. Since $\rho pm = o(1)$,

$$\Pr[s_i = 0 \wedge s_j = 0 \wedge s_k = 0] \geq 1 - \Pr[s_i = 1] - \Pr[s_j = 1] - \Pr[s_k = 1] \geq 1 - 3pm\rho = \Omega(1)$$

Hence, by Corollary 113, with a number of samples as stated in the theorem,

$$\|\hat{\text{PMIT}}_{S_a,S_b,S_c} - \text{PMIT}_{S_a,S_b,S_c}\|_{\{1,2\},\{3\}} \lesssim p^{5/2}\rho^{5/2}\sqrt{mn}^2 \tag{3.3.42}$$

as well.

With that, invoking Theorem 10 (with $\|E\|_{\{1,2\},\{3\}}$ taking into account both the E_L term above, and the above error due to sampling), the output of Algorithm 12 will produce vectors v_i, $i \in [m]$, s.t. v_i is $O(\eta')$-close to $\left(\frac{\rho}{1-\rho}\right)^{1/3}(1 - \exp(-W_i))$, for

$$\eta' \lesssim \max(\|Q_a\|, \|Q_b\|, \|Q_c\|)^{1/2} \cdot \left(\tau^{3/2} + \sigma^{-3/2}(\|E_L\|_{\{1,2\},\{3\}} + \|\hat{\text{PMIT}}_{S_a,S_b,S_c} - \text{PMIT}_{S_a,S_b,S_c}\|_{\{1,2\},\{3\}})\right)$$

where $\sigma = \min(\sigma_{\min}(Q_a), \sigma_{\min}(Q_b), \sigma_{\min}(Q_c))$.

Plugging in the estimates from (3.3.40), (3.3.41), (3.3.42) as well as $\tau = O(\rho^{2/3}\log n)$, we get:

$$\begin{aligned}
\max(\|Q_a\|, \|Q_b\|, \|Q_c\|)^{1/2}\tau^{3/2} &\lesssim \sqrt{mnp^2\rho^{2/3}}(\rho^{2/3}\log n)^{3/2} = \rho^{4/3}\sqrt{mn}p\log^{3/2} n \\
\sigma^{-3/2}\|E_L\|_{\{1,2\},\{3\}} &\lesssim (\frac{1}{\rho np})^{3/2}\|E_L\|_{\{1,2\},\{3\}} \lesssim \rho^{4/3}\sqrt{mn}p\log^{3/2} n \\
\sigma^{-3/2}\|\hat{\text{PMIT}}_{S_a,S_b,S_c} - \text{PMIT}_{S_a,S_b,S_c}\|_{\{1,2\},\{3\}} &\lesssim \rho^{4/3}\sqrt{mn}p\log^{3/2} n
\end{aligned}$$

which implies the vectors $(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ are $O(\eta')$-close to $\left(\frac{\rho}{1-\rho}\right)^{1/3}(1 - \exp(W_i))$, for all $i \in [m]$.

However, this directly implies that $\left(\frac{1-\rho}{\rho}\right)^{1/3}(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ are $O(\eta'/\rho^{1/3})$-close to $1 - \exp(W_i)$, $i \in [m]$, which in turn

implies $(\tilde{a}_i, \tilde{b}_i, \tilde{c}_i)$ are $O(\eta'/\rho^{1/3})$ close to $\exp(W_i)$.

This implies the statement of the Lemma.

□

Given that, we prove the main theorem. The main issue will be to ensure that taking log of the values

Proof of Theorem 7. By Proposition 110, the vectors $Y_i, i \in [m]$ in Algorithm 17 are $O(\eta\sqrt{np})$-close to $\exp(-W_i)$ with $\eta = \tilde{O}\left(\sqrt{mp}\rho\right)$. Let $(Y_i')_j = (Y_i)_j$ if $(Y_i)_j > \exp(-\nu_u)$ and otherwise $(Y_i)_j' = \exp(-\nu_u)$.

Then we have that $\|Y_i' - W_i\| \leq \|Y_i - W_i\|$. By the Lipschitzness of $\log(\cdot)$ in the region $[\nu_i, \infty]$

we have that

$$|(\widehat{W}_i)_j - (W_i)_j| = |\log(Y_i')_j - \log(W_i)_j| \lesssim |(Y_i)_j' - (W_i)_j|$$

It follows that

$$\|\widehat{W}_i - W_i\| = \|\log Y_i' - \log W_i\| \lesssim \|Y_i' - W_i\|$$

Therefore recalling $\|Y_i' - W_i\| \leq \|Y_i - W_i\| \leq O(\eta\sqrt{np})$ we complete the proof.

□

Proof of Theorem 8. The proof will follow the same outline as the proof of Theorem 7. The difference is that since we only have a guarantee on the spectral boundedness of the second and third-order term, we will need to bound the higher-order terms in a different manner. Given that we have no information on them in this scenario, we will simply bound them in the obvious manner. We proceed to formalize this.

The sample complexity is polynomial for the same reasons as in the proof of Theorem 7, so we will not worry about it here.

We only need to check the conditions for Algorithms 13 and 12 to work, so that we can apply Theorems 10 and 6.

Towards that, first we claim that we can write the PMI tensor for any equipartition S_a, S_b, S_c of $[n]$ as

$$\begin{aligned}
&\mathrm{PMIT}_{S_a,S_b,S_c} = \\
&\frac{\rho}{1-\rho}\sum_{k\in[m]} F_{k,S_a} \otimes F_{k,S_b} \otimes F_{k,S_c} - \left(\frac{1}{2}\left(\frac{\rho}{1-\rho}\right)^2\right)\sum_{k\in[m]} G_{k,S_a} \otimes G_{k,S_b} \otimes G_{k,S_c} \\
&+\left(\frac{1}{3}\left(\frac{\rho}{1-\rho}\right)^3\right)\sum_{k\in[m]} H_{k,S_a} \otimes H_{k,S_b} \otimes H_{k,S_c} + E
\end{aligned} \tag{3.3.43}$$

where $\|E\|_{\{1,2\},\{3\}} \leq \rho^4 m(np)^{3/2}$. Towards achieving this, first we claim that Proposition 94 implies that for any subsets S_a, S_b, S_c,

$$\left\| \sum_{k=1}^{m} (1-\exp(-lW_k))_{S_a} \otimes (1-\exp(-lW_k))_{S_b} \otimes (1-\exp(-lW_k))_{S_c} \right\|_{\{1,2\},\{3\}} \leq m(np)^{3/2}$$

Indeed, if we put $\gamma_k = (1-\exp(-lW_k))_{S_a}$, $\delta_k = (1-\exp(-lW_k))_{S_b}$, $\theta_k = (1-\exp(-lW_k))_{S_c}$, then we have $\|\sum_k \gamma_k \gamma_k^\top\| \leq \sqrt{mnp}$, and similarly for δ_k. Since $\max_k \|\theta_k\| \leq (np)^{1/2}$, the claim immediately follows. Hence, (3.3.43) follows.

Next, let R_{S_a} be the matrix that has as columns the vectors $\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)^{1/3} (\tilde{P}_l)_{j,S_a}$, $l \in [2,L]$, $j \in [m]$ and A is the matrix that has as columns the vectors $\left(\frac{\rho}{1-\rho}\right)^{1/3} (\tilde{P}_1)_{j,S_a}$ for $j \in [m]$ for some $L = O(\mathrm{poly}(n))$, similarly as in the proof of Theorem 7.

We claim that $R_{S_a} R_{S_a}^\top$ is τ spectrally bounded by $\rho FF^\top$.

Indeed, for any $l > 2$, we have $\|\left(\frac{1}{l}\left(\frac{\rho}{1-\rho}\right)^l\right)^{1/3} \tilde{P}_l\| \lesssim \rho^{l/3}\|\tilde{P}_l\| \lesssim \rho^{l/3}\sqrt{mn}p$ Hence,

$$\begin{aligned} R_{S_a} R_{S_a}^\top &\lesssim \rho^{2/3} GG^\top + \rho^{4/3} HH^\top + \rho^2 LL^\top + \sum_{l \geq 4} \rho^{2l/3} mnp^2 \\ &\lesssim 3\rho^{2/3}\tau(FF^\top + \sigma_{\min}(FF^\top)) + \rho^{8/3} mnp^2 \lesssim \rho^{2/3}\tau(FF^\top + \sigma_{\min}(FF^\top)) \end{aligned} \tag{3.3.44}$$

where the first inequality holds since $HH^\top, GG^\top, LL^\top$ are τ-spectrally bounded bounded by F and the second since $\sigma_{\min}(FF^\top) \gtrsim np$ and $\tau \geq 1$. Let $\tau' = \rho^{2/3}\tau$. Since we are assuming the matrix F is $O(1)$-incoherent, we can apply Theorem 16, and claim the output of Algorithm 13 are matrices Q_a, Q_b, Q_c which are τ-approximate whitening matrices for A, B, C respectively.

By Lemma 109, we have again

$$\sigma_{\max}(Q_a) \leq \frac{1}{1-\tau'} \lesssim (1+\tau')\|AA^\top\|, \sigma_{\min}(Q_a) \geq \frac{1}{1+\tau'} \gtrsim (1-\tau')\sigma_{\min}(AA^\top)$$

Then, applying Theorem 10, we get that we recover vectors $(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ are $O(\eta')$-close to $\left(\frac{\rho}{1-\rho}\right)^{1/3}(1-\exp(W_i))$, for all $i \in [m]$. for

$$\eta' \lesssim \max(\|Q_a\|, \|Q_b\|, \|Q_c\|)^{1/2} \cdot \left(\tau'^{3/2} + \sigma^{-3/2}\|E\|_{\{1,2\},\{3\}}\right)$$

Recall that $\tau' = \rho^{2/3}\tau$, and $\|Q_a\| \leq \rho^{2/3}\sigma_{\max}(F) \lesssim \rho^{2/3}\sqrt{mn}p$ and $\|E\|_{\{1,2\},\{3\}} \leq \rho^4 m(np)^{3/2}$ and $\sigma \gtrsim \rho^{2/3}np$, we obtain that

$$\eta' \lesssim \rho^{1/3}\sqrt{mn}p\left((\tau\rho^{2/3})^{3/2} + \frac{\rho^4 m(np)^{3/2}}{(\rho^{2/3}np)^{3/2}}\right) \lesssim \rho^{4/3}\sqrt{mn}p\tau^{3/2}$$

where the last inequality holds since $\rho^3 m = o(1) = o(\tau)$.

However, this directly implies that $(\frac{1-\rho}{\rho})^{1/3}(\hat{a}_i, \hat{b}_i, \hat{c}_i)$ are $O(\eta'/\rho^{1/3}) = O(\eta)$-close to $1-\exp(-W_i)$, $i \in [m]$, which

in turn implies $(\tilde{a}_i, \tilde{b}_i, \tilde{c}_i)$ are $O(\eta)$ close to $\exp(-W_i)$.

Argument for recovering W_i from $\exp(-W_i)$ is then exactly the same as the one in Theorem 7.

□

3.3.12 Technical details: sample complexity and bias of the PMI estimator

Finally, we consider the issue of sample complexity. The estimator we will use for the PMI matrix will simply be the plug-in estimator, namely:

$$\hat{\mathrm{PMI}}_{i,j} = \log \frac{\hat{\Pr}[s_i = 0 \wedge s_j = 0]}{\hat{\Pr}[s_i = 0]\hat{\Pr}[s_j = 0]} \tag{3.3.45}$$

Notice that this estimator is biased, but as the number of samples grows, the bias tends to zero. Formally, we can show:

Lemma 111. *If the number of samples N satisfies*

$$N \geq \frac{1}{\min_{i \neq j}\{\Pr[s_i = 0 \wedge s_j = 0\}} \frac{1}{\delta^2} \log m$$

with high probability $|\hat{\mathrm{PMI}}_{i,j} - \mathrm{PMI}_{i,j}| \leq \delta, \forall i \neq j$.

Proof. Denoting $\Delta_{i,j} = \hat{\Pr}[s_i = 0 \wedge s_j = 0] - \Pr[s_i = 0 \wedge s_j = 0]$ and $\Delta_i = \hat{\Pr}[s_i = 0] - \Pr[s_i = 0]$, we get that

$$\begin{aligned}\hat{\mathrm{PMI}}_{i,j} &= \log \frac{\hat{\Pr}[s_i = 0 \wedge s_j = 0]}{\hat{\Pr}[s_i = 0]\hat{\Pr}[s_j = 0]} = \log \frac{\Pr[s_i = 0 \wedge s_j = 0] + \Delta_{i,j}}{(\Pr[s_i = 0] + \Delta_i)(\Pr[s_j = 0] + \Delta_j)} \\ &= \mathrm{PMI}_{i,j} + \log\left(1 + \frac{\Delta_{i,j}}{\Pr[s_i = 0 \wedge s_j = 0]}\right) - \log\left(1 + \frac{\Delta_i}{\Pr[s_i = 0]}\right) - \log\left(1 + \frac{\Delta_j}{\Pr[s_j = 0]}\right)\end{aligned}$$

Furthermore, we have that $\frac{2x}{2+x} \leq \log(1+x) \leq \frac{x}{\sqrt{x+1}}$, for $x \geq 0$, which implies that when $x \leq 1$, $\frac{2}{3}x \leq \log(1+x) \leq x$. From this it follows that if

$$\max\left(\max_{i,j} \frac{\Delta_{i,j}}{\Pr[s_i = 0 \wedge s_j = 0]}, \max_i \frac{\Delta_i}{\Pr[s_i = 0]}\right) \leq \delta$$

we have

$$\mathrm{PMI}_{i,j} - \frac{\delta}{3} \leq \hat{\mathrm{PMI}}_{i,j} \leq \mathrm{PMI}_{i,j} + \delta$$

Note that it suffices to show that if $N > \frac{1}{1-4p_{\max}m\rho_{\max}} \frac{1}{\delta^2} \log m$, we have

$$\Pr\left[\frac{\Delta_i}{\Pr[s_i = 0]} > (1+\delta) \vee \frac{\Delta_i}{\Pr[s_i = 0]} < (1-\delta)\right] \leq \exp(-\log^2 m) \tag{3.3.46}$$

and

$$\Pr\left[\frac{\Delta_{i,j}}{\Pr[s_i = 0 \wedge s_j = 0]} > (1+\delta) \vee \frac{\Delta_{i,j}}{\Pr[s_i = 0 \wedge s_j = 0]} < (1-\delta)\right] \leq \exp(-\log^2 m) \tag{3.3.47}$$

since this implies

$$\max\left(\max_{i,j} \frac{\Delta_{i,j}}{\Pr[s_i = 0 \wedge s_j = 0]}, \max_i \frac{\Delta_i}{\Pr[s_i = 0]}\right) \leq \delta$$

with high probability by a simple union bound.

Both (3.3.46) and (3.3.47) will follow by a Chernoff bound.

Indeed, consider (3.3.46) first. We have by Chernoff

$$\Pr\left[\Delta_i > \left(1 + \sqrt{\frac{\log N}{N \Pr[s_i = 0]}}\right)\Pr[s_i = 0]\right] \leq \exp(-\log^2 N)$$

Hence, if $N > \frac{1}{\Pr[s_i=0]}\frac{1}{\delta^2}\log m$, we get that $1 - \delta \leq \frac{\Delta_i}{\Pr[s_i=0]} \leq 1 + \delta$ with probability at least $1 - \exp(\log^2 m)$.

The proof of (3.3.47) is analogous – the only difference being that the requirement is that $N > \frac{1}{\Pr[s_i=0]}\frac{1}{\delta^2}\log m$ which gives the statement of the lemma.

□

Virtually the same proof as above shows that:

Lemma 112. *If the number of samples N satisfies*

$$N \geq \frac{1}{\min_{i\neq j\neq k}\{\Pr[s_i = 0 \wedge s_j = 0 \wedge s_k = 0\}} \frac{1}{\delta^2}\log m$$

with high probability $|\hat{\text{PMIT}}_{i,j,k} - \text{PMIT}_{i,j,k}| \leq \delta, \forall i \neq j \neq k$.

As an immediate corollary, we get:

Corollary 113. *If the number of samples N satisfies*

$$N \geq \frac{1}{\min_{i\neq j\neq k}\{\Pr[s_i = 0 \wedge s_j = 0 \wedge s_k = 0\}} \frac{1}{\delta^2}\log m$$

$$N \geq \frac{1}{\min_{i\neq j\neq k}\{\Pr[s_i = 0 \wedge s_j = 0 \wedge s_k = 0\}} \frac{1}{\delta^2}\log m$$

with high probability for any equipartition S_a, S_b, S_c

$$\|\hat{\text{PMIT}}_{S_a,S_b,S_c} - \text{PMIT}_{S_a,S_b,S_c}\|_{\{1,2\},\{3\}} \lesssim n^3\delta$$

3.3.13 Technical details: matrix perturbation toolbox

In this section we discuss standard matrix perturbation inequalities. Many results in this section can be found in Stewart and Sun (Stewart, 1977). Given $\widehat{A} = A + E$, the perturbation in individual singular values can be bounded by Weyl's theorem:

Theorem 17 (Weyl's theorem). *Given $\widehat{A} = A + E$, we know $\sigma_k(A) - \|E\| \leq \sigma_k(\widehat{A}) \leq \sigma_k(A) + \|E\|$.*

For singular vectors, the perturbation is bounded by Wedin's Theorem:

Lemma 114 (Wedin's theorem; Theorem 4.1, p.260 in (Stewart, 1990)). *Given matrices $A, E \in \mathbb{R}^{m\times n}$ with $m \geq n$. Let A have the singular value decomposition*

$$A = [U_1, U_2, U_3] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \\ 0 & 0 \end{bmatrix} [V_1, V_2]^\top.$$

Let $\widehat{A} = A + E$, with analogous singular value decomposition. Let Φ be the matrix of canonical angles between the column span of U_1 and that of $\widehat{U}_1$, and Θ be the matrix of canonical angles between the column span of V_1 and that of $\widehat{V}_1$. Suppose that there exists a δ such that

$$\min_{i,j} |[\Sigma_1]_{i,i} - [\Sigma_2]_{j,j}| > \delta, \quad \textit{and} \quad \min_{i,i} |[\Sigma_1]_{i,i}| > \delta,$$

then

$$\|\sin(\Phi)\|^2 + \|\sin(\Theta)\|^2 \leq 2\frac{\|E\|^2}{\delta^2}.$$

Perturbation bound for pseudo-inverse When we have a lowerbound on $\sigma_{min}(A)$, it is easy to get bounds for the perturbation of pseudoinverse.

Theorem 18 (Theorem 3.4 in (Stewart, 1977)). *Consider the perturbation of a matrix $A \in \mathbb{R}^{m\times n}$: $B = A + E$. Assume that $rank(A) = rank(B) = n$, then*

$$\|B^\dagger - A^\dagger\| \leq \sqrt{2}\|A^\dagger\|\|B^\dagger\|\|E\|.$$

Note that this theorem is not strong enough when the perturbation is only known to be τ-spectrally bounded in our definition.

3.4 Beyond linearity II: provable algorithms for noisy-OR using anchor symptoms

In this section, similar as in the case of topic models, we will design more efficient algorithms for noisy-OR which are based on solving a *non-linear* variant of non-negative matrix factorization, rather than tensor decomposition. The benefit will be two-fold: first, the structural assumptions will be completely transparent and verifiable – and in fact, satisfied on the QMR-DT dataset, which is the standard testbed for this problem; second, the algorithms will be computationally substantially more efficient – as was the case for topic models. In fact, the only obstacle to getting completely practical algorithms will be sample complexity: given the population values of the PMI matrix, the algorithm takes under an hour to learn the actual QMR-DT network; unfortunately, the sample complexity for estimating the entries of the PMI matrix to a sufficient accuracy will still be rather prohibitive in practice.

The main contribution of this section is to combine the approach of linearizing the PMI matrix in Section 3.3 with a fast algorithm for symmetric nonnegative matrix factorization (NMF), which provably works under a new structural assumption about noisy-or networks that we call *sequential 2-anchor condition*(see (A2) in Section 3.4.0.1). This lets us design a new algorithm whose running time is n^3, improving both on the runtime of Section 3.3 and the n^4 runtime required for quartet learning in (Jernite et al., 2013). Our algorithm actually can be extended to a certain nonlinear variant of symmetric NMF, which is defined in the following section.

3.4.0.1 Overview of assumptions and approach: a meta-algorithm for non-linear Symmetric NMF

We proceed now to describing our approach using algorithm for symmetric NMF. First, we set up some notation and nomenclature that will be used throughout the paper. We introduce an analogue of an anchor word in (Arora et al., 2013b):

Anchor row: An anchor row of a matrix X is a row that has a singleton support:

Definition (Anchor row)**.** An anchor row of the matrix $X \in \mathbb{R}^{n \times m}$ is a row of X with only one non-zero entry.

The anchor word assumption in (Arora et al., 2013b) in this language requires that for every column index j, there exists an anchor row i such that $\mathrm{supp}(X^i) = \{j\}$. Besides generalizing to non-linear situation, one of the features of this paper is to weaken the anchor word assumption. (See Section 3.4.3) and Assumption A2 for details.

Our approach will solve a generalization of the symmetric NMF problem we call sym-NMF. In sym-NMF, the observed matrix is the sum of a nonlinear function applied to each rank-1 component:

Definition (Non-linear sym-NMF)**.** Let $f : \mathbb{R}^{n \times n}_{\geq 0} \to \mathbb{R}^{n \times n}_{\geq 0}$ be some *known* function that satisfies *zero preservation* in the sense that for all (i, j), $f(B)_{ij} = 0$ if and only if $B_{ij} = 0$. Let $X_1, \ldots, X_m$ be unknown vectors in $\mathbb{R}^n_{\geq 0}$. Given the

matrix $A \in \mathbb{R}^{n\times n}$ of the form

$$A = \sum_{j=1}^{m} f\left(X_j X_j^\top\right), \tag{3.4.1}$$

the problem of *non-linear sym-NMF* asks to recover the vectors X_j's.

The standard NMF problem of course corresponds to the sub-case where f is the identity mapping from $\mathbb{R}_{\geq 0}^{n\times n}$ to $\mathbb{R}_{\geq 0}^{n\times n}$. As for our original motivation – noisy-or networks, Proposition A.3 in (Arora et al., 2017b) implies that the problem of recovering the weight matrix W from PMI corresponds to the following choice of f in the non-linear sym-NMF problem. (Recall that in the definition , $X_j X_j^\top$ is always a symmetric nonnegative rank-1 matrix, thus we only to define f on such matrices.)

Lemma 115. *Suppose $f : \mathbb{R}_{\geq 0}^{n\times n} \to \mathbb{R}_{\geq 0}^{n\times n}$ satisfies that for any symmetric nonnegative rank-1 Z of the form $Z = zz^\top$, we have*

$$f(Z) = \sum_{k=1}^{\infty} \beta_k (1 - (1-z)^{\odot k}), \quad \text{where } \beta_k = (-1)^{k+1} \frac{1}{k} \left(\frac{\rho}{1-\rho}\right)^k$$

Then, the PMI matrix PMI *for the noisy-or problem satisfies that*

$$\text{PMI} = \sum_{j=1}^{m} f(F_j F_j^\top) + D, \tag{3.4.2}$$

where $F_j = 1 - \exp(-W_j)$, and D is a diagonal matrix.

In words, except for the diagonal entries, PMI can be written as sum of non-linear function of the rank matrics $F_j F_j^\top$.

We design a meta-algorithm for sym-NMF, described in Algorithm 14, which consists of three generic steps. For each one, we subsequently provide algorithms, and conditions under which they succeed. (Section 3.4.1 for Step 1, Section 3.4.2 for Step 2 and Section 3.4.3 for Step 3.)

3.4.1 Step 1: anchor rows discovery

In the first step of the algorithm, we identify anchor rows by finding rows with minimal support. More precisely for row i, if there exists another row whose support is strictly within the support of row i, then it is not an anchor row. (Algorithm 15) We show this simple strategy works under the following assumption:

Assumption A1. *[Missing triangle condition] For any two columns j, ℓ with overlapping supports, there exists row i, k with disjoint supports such that $i \in \text{supp}(X_j) \setminus \text{supp}(X_\ell)$ and $k \in \text{supp}(X_\ell) \setminus \text{supp}(X_j)$.*

[8] The number j is not identifiable since the the potential solutions are equivalent up to permutation of the columns.

Algorithm 14 Meta-algorithm for non-linear NMF

Given: matrix function f and matrix $A = \sum_{j=1}^{m} f(X_j X_j^\top)$ as in definition .
Output: vectors $X_1, \ldots, X_m$ (up to permutation).

Repeat until all of $X_1, \ldots, X_m$ are found

Step 1 Anchor rows discovery: Find the indices i, k such that row i and row k are two anchor rows of the matrix X that share same support (which is a singleton by definition).

Step 2 Column recovery: Given rows i, k both with support $\{j\}$[8] Recover the column X_j.

Step 3 Peeling-off: Remove the contribution of X_j from A (by subtracting $f(X_j X_j^\top)$ from A)

Algorithm 15 Anchor detection

Given: A matrix $A = \sum_{j=1}^{m} f(X_j X_j^\top)$ and an row i.
Outputs No, if $\exists k \neq i \in [n]$ such that $i \in \text{supp}(A^k)$ and $\text{supp}(A^i) \not\subseteq \text{supp}(A^k)$; otherwise, Yes

The reason we call condition (A1) the *missing triangle condition* is apparent from its graphical representation, as shown on the left part of Figure 3.1; the blue ellipsoids in the figure denote the supports of columns j and l, and m is used to indicate an arbitrary column; the dashed lines indicate that m doesn't belong to support of row i or k, and a full line denotes it does. The "missing triangles" refers to the fact that for every column m, it doesn't belong to the support of at least one of the rows i or k.

In the contexts of noisy-or networks and QMR-DT dataset , this means that for any two diseases that share at least one symptom, there are two symptoms in the support of these two diseases respectively, but outside of the intersection, that don't share a common disease. What this condition intuitively ensures is that if a symptom isn't an anchor symptom, when we try to treat it as an anchor symptom and "subtract" its influence in the PMI matrix, we will see negative values.

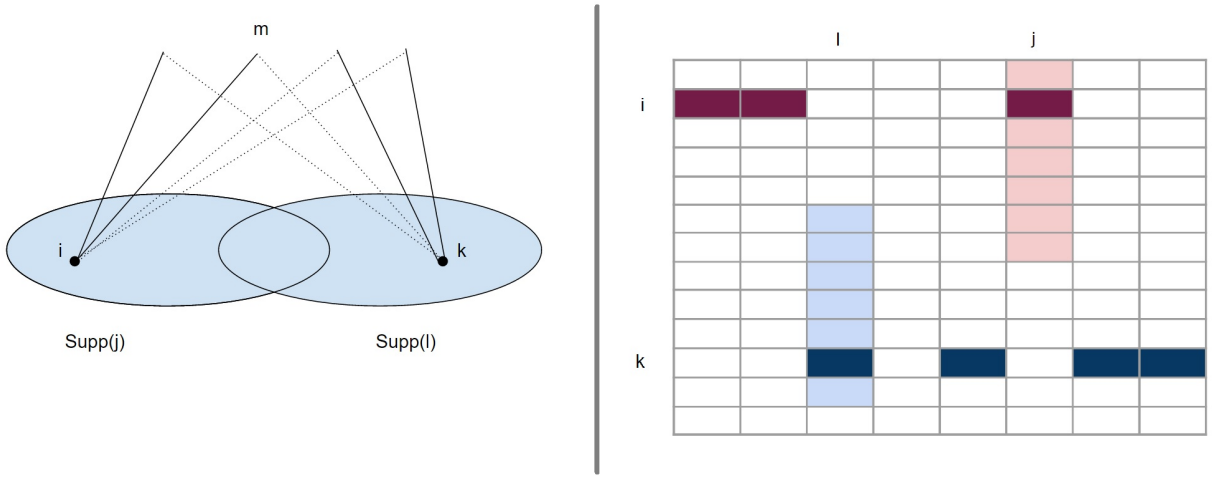
More formally, the following guarantee holds:

Lemma 116. *Under the assumption A1, algorithm 15 outputs Yes iff coordinate i is an anchor-row.*

Proof. Assume first that i is an anchor row for some column j. Because f is zero-preserving, $\text{supp}(A) = \text{supp}(XX^\top)$. This implies that $\forall k, \text{supp}(A^k) = \sum_{j=1}^{m} \text{supp}(X_{k,j} X^j)$. But, since i is an anchor row for j, the only rows k for which $i \in \text{supp}(A^k)$ will have also have $X_{k,j} > 0$. But this clearly implies $\text{supp}(A^j) \subseteq \text{supp}(A^k)$, so in this case, Algorithm 15 will output Yes, as we want.

Next, assume i is not an anchor row, namely exist two columns j, l, s.t. $X_{i,j} > 0, X_{i,l} > 0$. By (A1), there exist also two rows a, k with disjoint supports outside of the intersection of $\text{supp}(X_j) \cap \text{supp}(X_l)$, but in the support of X_j and X_l respectively. Then, we claim that $\text{supp}(A^i) \not\subseteq \text{supp}(A^a)$. More precisely, we will show that $k \in \text{supp}(A^i)$ but $k \notin \text{supp}(A^a)$. Indeed, $k \in \text{supp}(A^i)$ follows since $a \in \text{supp}(X_j)$ and $X_{i,j} > 0$. On the other hand, $k \notin \text{supp}(A^a)$ follows since for any $s \in \text{supp}(X^a)$, $k \notin \text{supp}(A^s)$, so $k \notin \text{supp}(A^a)$.

Figure 3.1: Two graphical representations of condition (A1), on the left in Venn diagram form, on the right in matrix form. The blue ellipsoids on the left denote the supports of columns j and l, and m is an arbitrary column. Dashed lines indicate m doesn't belong to support of row i or k, and full denotes it does. "Missing triangles" refers to the fact that for every column m, it doesn't belong to the support of at least on of the rows i or k. On the right, the same condition is represented in matrix form, in which colored-in squares denote non-zero elements.



□

3.4.2 Step 2: column recovery

The key challenge of the algorithm is the second step where we aim to recover some column of the matrix X by leveraging the information of the anchor rows. This step needs to be specialized to the choice of function of f. In the following subsection, we present an algorithm for case of *invertible* and *decomposable* functions f, which sheds light on the general idea and is of independent interests. Later in Section 3.4.4, we will design an substantially more advanced algorithm that handles a class of function f, using which we can learn the noisy-or networks.

Definition. We call a mapping $f : \mathbb{R}_{\geq 0}^{n\times n} \to \mathbb{R}_{\geq 0}^{n\times n}$ *decomposable* if f applies on each coordinate individually, or concretely, if $f(Z)$ can be written as $\{f_{ij}(Z_{ij})\}_{(i,j)\in[n]\times[n]}$ for some single-value functions f_{ij}'s. We call it *invertible* if f is a one-to-one map.

Lemma 117. *Suppose that f is decomposable and invertible and row i is an anchor row of X. Then Algorithm 16 returns a vector v that is equal to one of the columns of X.*

Proof. Suppose row i of X has support $\text{supp}(X^i) = \{s\}$ Then we claim that $v = X_s$. By the definition of A, that the support of X^i is $\{s\}$ and f preserves zero entry wise, we have that $A^i = f(X_s X_s^\top)^i = [f_{i1}(X_{is}X_{1s}), f_{i2}(X_{is}X_{2s}), \ldots, f_{n1}(X_{is}X_{ns})]$ where the last equality uses the fact that f is decomposable. Now inverting the function f we obtain that $w =$

Algorithm 16 Recovery algorithm for decomposable invertible function f

Inputs: A decomposable and invertible function $f = \{f_{ij}\}$ and invertible A matrix $A = \sum_{j=1}^{m} f(X_j X_j^\top)$, an anchor row i.

Outputs: A vector v (one of the columns of X)

1. Let $w = [f_{i1}^{-1}(A_{i1}), f_{i2}^{-1}(A_{i2}), \ldots, f_{in}^{-1}(A_{in})]$
2. Let $v_i = \sqrt{w_i}$ and $\forall \ell \neq i$, let $v_\ell = w_\ell / v_i$.
3. Return v

$[f_{i1}^{-1}(A_{i1}), f_{i2}^{-1}(A_{i2}), \ldots, f_{in}^{-1}(A_{in})] = [X_{is}X_{1s}, X_{is}X_{2s}, \ldots, X_{is}X_{ns}]$. Then we can see that $X_{is} = \sqrt{w_i}$ and $X_{\ell s} = w_\ell / X_{is} = w_\ell / v_i$. □

3.4.3 Iterative peeling-off

Assumption A2. *[*Sequential 2-anchor condition*]: After removal of any subset U of columns from X, the remaining matrix $X_{[m]\setminus U}$ still has at least one pair of anchor rows with the same support.*

Note that if we needed a single anchor row, this assumption would be weaker than the "anchor word" assumption in (Arora et al., 2013b): in this paper's language, this assumption requires that for every column j, there exists an anchor row i such that $\mathrm{supp}(X^i) = \{j\}$. Thus it requires at least m anchors in the matrix X, each with different support. Here we are satisfied with a pair of anchors for a *single* column j, but we require more anchor row to appear after the columns with identified anchors are "peeled off". In fact, we know that the QMR-DT dataset satisfies our condition, but doesn't satisfy the anchor word assumption in (Arora et al., 2013b).

The assumption that we have a *pair* of anchor rows for a particular column is necessary for our noisy-or algorithm – more concretely, for Step 2 of recovering a column X_j. For the simpler case of decomposable and invertible f, in fact even a single anchor row would suffice – hence our assumption is a strict relaxation of the "anchor word" assumption.

For computational efficiency, we would also like to remove the columns of X in stages: in each stage, we discover multiple anchor row pairs, and then remove all of the corresponding columns simultaneously. We call the number of stages to peel off all of the columns in a greedy fashion **greedy peeling depth**.

It's obvious that assuming Step 1 and Step 2 in the Algorithm 14 are correct, under assumption A2, we can inductively remove all the columns from X until recovering all of the columns. Applying this to the situation of decomposable and invertible function f for which we have shown recovery algorithm in the Section 3.4.2, we obtain the following result:

Theorem 19 ((Arora et al., 2017a))**.** *Under Assumption A1 and A2, given matrix A of the form (3.4.1), we can recover the vector $X_1, \ldots, X_m$ up to permutation in polynomial time. In addition, if the greedy peeling depth is T, then the runtime is at most $O(Tn^3)$.*

Proof of Theorem 19. The proof of correctness is immediate – only the runtime requires proof. Consider any stage in the peeling off process. All anchor rows at that stage can be discovered in time $O(n^3)$: we iterate over all rows i, and implement Algorithm 15 in the trivial manner by looping over all other rows k, checking whether their supports agree in the trivial manner. Subsequently, we recover the columns corresponding to the anchor rows by running Algorithm 16 in time $O(n)$ per anchor-row – hence $O(n^2)$ in total. Finally, a column can be removed in time $O(n^2)$, so we can remove all the columns from a particular stage in time $O(n^3)$. Overall, the runtime is $O(n^3)$ per stage, which implies the theorem.

□

3.4.4 Main result: learning noisy-or networks via non-linear sym-NMF

In this section, we show how meta-algorithm for sym-NMF can be adapted to learn noisy-or networks. The main difficulty here is how to recover the network parameters from the anchor symptoms. In order to do that we will also keep track of the first and second order moments $p(\bar{s}_i)$ and $p(\bar{s}_i, \bar{s}_k)$ for all pairs of symptoms i, k. We give the main algorithm in Algorithm 17 and prove the following guarantees.

Theorem 20 ((Arora et al., 2017a))**.** *Suppose assumption A1 and A2 hold, with greedy peeling depth T. Moreover, suppose in Algorithm 17 we are given the exact PMI matrix* PMI *and in addition to the first and second moments of the symptoms s. Then algorithm 17 recovers the weight matrix exactly in time $O(Tn^3)$.*

Algorithm 17 is very similar to Algorithm 14. The main difference is how we recover the weights of the network from anchor symptoms. We describe this step (Algorithm 18) in Section 3.4.5. In Step 3 we also need to maintain the first and second moments, which appears in Algorithm 20 and Section 3.4.5.1. Finally in Section 3.4.5.2 we describe how to make Theorem 20 robust to noise.

Algorithm 17 Learning noisy-or networks

Inputs: Empirical PMI matrix $\hat{\text{PMI}}$, empirical estimates of $\hat{p}(\bar{s}_i, \bar{s}_j)$, $\hat{p}(\bar{s}_i)$.
Outputs: Estimate $\widehat{W}$ of weight matrix.
Repeat until all of $W_1, \ldots, W_m$'s are found

Step 1 Anchor rows discovery: Find the indices i, k such that row i and row k are two anchor rows of the matrix X that share same support (which is a singleton by definition).

Step 2 Column recovery: Suppose the supports of row i and k are both $\{j\}$[9] Recover the column W_j using algorithm Algorithm 18.

Step 3 Peeling-off: Remove the contribution of disease j from PMI matrix and adjust the empirical moments using Algorithm 20.

[9] As noted before, the number j is not identifiable since the the potential solutions are equivalent up to permutation of the columns.

3.4.5 Column recovery algorithm for learning noisy-or

Suppose we are trying to recover disease j with two anchor symptoms i and k. The basic observation is that when we observe an anchor symptom, we know the disease must be on. With more work, we can additionally relate the observed pairwise correlations of various symptoms with parameters of the disease. Formalizing this, we can calculate the parameters W_j by solving quadratic equations, with coefficients calculated from empirical marginals of the symptoms. More precisely we have

Lemma 118. *Let i, k be anchor symptoms for disease j such that $W_{i,j} \neq W_{k,j}$, let y, z two arbitrary symptoms. Let a, b be 2-dimensional vectors, defined as*

$$a_1 = \frac{p(s_y = 0|s_i = 0)}{p(s_y = 0|s_i = 1)}, b_2 = \frac{p(s_z = 0|s_i = 0)}{p(s_z = 0|s_i = 1)}, b_1 = \frac{p(s_y = 0|s_k = 0)}{p(s_y = 0|s_k = 1)}, b_2 = \frac{p(s_z = 0|s_k = 0)}{p(s_z = 0|s_k = 1)} \tag{3.4.3}$$

Furthermore, for a vector v, let $R_{-1}v$ be the projection of v to the direction orthogonal to $\mathbf{1}$. (Note that $R_{-1}v = v_1 - v_2$.)

Then, if p_1, p_2 are the solutions to the following quadratic equation in χ:

$$\frac{R_{-1}b}{R_{-1}a}(1-\chi)(1-\rho) = \frac{p(s_i = 0|s_l = 0)}{p(s_i = 0|s_l = 1)}\chi(\chi\rho + 1 - \rho) \tag{3.4.4}$$

exactly one of p_1, p_2 is positive and equals $1 - \exp(-W_{i,j})$.

Using similar ideas we can recover all the other weights for disease j:

Lemma 119. *Let i be an anchor symptom for disease j. Let a' be the vector defined $a'_z = \frac{p(s_z=0|s_i=1)}{p(s_z=0|s_i=0)}, z \in [n]$ and $r = \frac{(1-\exp(-W_{i,j}))\rho}{(1-\exp(-W_{i,j}))\rho+(1-\rho)}$. For all $z \neq i$, we have $W_{z,j} = -\log\left(\frac{a'_z-1}{a'_z-r}\right)$*

The proofs of these lemmas are somewhat technical, but essentially involve careful manipulation of the expressions for marginals in the noisy-or model. The algorithm to find a disease is a straightforward implementations of the statements of the lemmas as outlined in Algorithm 18

Let us now prove the above lemmas:

Proof of Lemma 118. Denote for notational convenience $p_x = 1-\exp(-W_{x,j})$ and $q_x = p(d_j = 1|s_x = 0)$. Furthermore, let $r_x = p(s_x = 0|d_j = 0)$. We claim that the lemma will follow easily from the following statements:

$$\frac{p(s_y = 0|s_i = 0)}{p(s_y = 0|s_i = 1)} = \frac{(1 - q_i)}{p_y} + q_i \tag{3.4.5}$$

$$\frac{p(s_k = 0|s_i = 0)}{p(s_k = 0|s_i = 1)} = \frac{1 - q_i}{p_k} + q_i \tag{3.4.6}$$

Algorithm 18 Reconstruct row W_j for disease j

Inputs: Empirical PMI matrix $\hat{\text{PMI}}$, empirical estimates of $\hat{p}$ all pairwise marginals, disease j to reconstruct, two tentative anchor symptoms i, k.
Outputs: Estimate of weight vector for disease $\widehat{W_j}$.

1. Pick two other diseases y, z randomly.
2. Let a, b be 2-dimensional vectors, defined as in equation (3.4.3)
3. Let $R_{-1}v$ be the projection of v to the direction orthogonal to $\mathbf{1}$. Note that $R_{-1}a = a_1 - a_2$, $R_{-1}b = b_1 - b_2$. Let q_1 be the (only) positive solution of equation (3.4.4) in χ
4. Set $W_{i,j} = -\log(1 - q_1)$. Let a' be the n-dimensional vector defined $a'_t = \frac{p(s_t=0|s_i=1)}{p(s_t=0|s_i=0)}, t \in [n]$.
5. Let $r = \frac{(1-\exp(-W_{i,j}))\rho}{(1-\exp(-W_{i,j}))\rho+(1-\rho)}$. For all $t \neq i$, set $W_{t,j} = -\log(\frac{a'_t-1}{a'_t-r})$

Let us show that first. Indeed, if a, b are the 2-dimensional vectors, as specified in the Lemma statement, (3.4.5) implies that

$$R_{-1}a = (1 - q_i)R_{-1}\frac{1}{p_y}$$

$$R_{-1}b = (1 - q_k)R_{-1}\frac{1}{p_y}$$

from which it follows that

$$\frac{R_{-1}a}{R_{-1}b} = \frac{1 - q_i}{1 - q_k} \tag{3.4.7}$$

Furthermore, note that

$$q_k = p(d_j = 1|s_k = 0) = \frac{p(d_j = 1)p(s_k = 0|d_j = 1)}{p(s_k = 0)} = \frac{\rho p_k}{\rho p_k + (1 - \rho)} \tag{3.4.8}$$

Putting together (3.4.6) and (3.4.8), we get

$$\frac{p(s_k = 0|s_i = 0)}{p(s_k = 0|s_i = 1)} - 1 = (1 - q_i)\left(1 - \frac{1}{p_k}\right)$$

and

$$\frac{1 - \rho}{\rho p_k + (1 - \rho)} = 1 - q_k$$

Diving the left and right-hand sides respectively, we get

$$\frac{\frac{p(s_k=0|s_i=0)}{p(s_k=0|s_i=1)} - 1}{\frac{1-\rho}{\rho p_k+(1-\rho)}} = \frac{1 - q_i}{1 - q_k}\left(1 - \frac{1}{p_k}\right) \tag{3.4.9}$$

Putting (3.4.7) and (3.4.9) together, we get the statement of the lemma.

Given that, we proceed to showing (3.4.5) and (3.4.6).

Consider (3.4.5) first. We will in fact show that

$$p(s_y = 0|s_i = 1) = p_y r_y \tag{3.4.10}$$

and

$$p(s_y = 0|s_i = 0) = (1 - q_i)r_y + q_i p_y r_y \tag{3.4.11}$$

First, consider the (3.4.10). As a notational convenience, let $d_{-j} = [1, m] \setminus \{j\}$ denote the set of diseases except for j.

By the law of total conditional probability, we have:

$$\begin{aligned}
p(s_y = 0|s_i = 1) &= \sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0, d_{-j} = D_{-j}, d_j = 1|s_i = 1) \\
&= \sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0|d_{-j} = D_{-j}, d_j = 1, s_i = 1)p(d_{-j} = D_{-j}, d_j = 1|s_i = 1) \\
&= \sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0|d_{-j} = D_{-j}, d_j = 1)p(d_j = 1|s_i = 1)p(d_{-j} = D_{-j}|s_i = 1, d_j = 1)
\end{aligned}$$

However, since i is an anchor symptom for disease j, $p(d_j = 1|s_i = 1) = 1$. Similarly, since $i \notin \text{supp}(l)$ for any $l \neq j$, we have

$$p(d_{-j} = D_{-j}|s_i = 1, d_j = 1) = p(d_{-j} = D_{-j}|d_j = 1) = p(d_{-j} = D_{-j})$$

. Finally, note that by (3.3.1), we have

$$p(s_y = 0|d_{-j} = D_{-j}, d_j = 1) = p(s_y = 0|d_{-j} = D_{-j}, d_j = 0)p_y$$

With this in mind, we have

$$\begin{aligned}
p(s_y = 0|s_i = 1) &= \sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0|d_{-j} = D_{-j}, d_g = 1)p(d_j = 1|s_i = 1)p(d_{-j} = D_{-j}|s_i = 1, d_j = 1) \\
&= p_y \sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0|d_{-j} = D_{-j}, d_j = 0)p(d_{-j} = D_{-j}) \\
&= p_y r_y
\end{aligned}$$

Similarly, for (3.4.11), we have

$$\begin{aligned}
p(s_y = 0 | s_i = 0) &= \sum_{d_j \in \{0,1\}, D_{-g} \in \{0,1\}^{m-1}} p(s_y = 0, d_{-j} = D_{-j}, d_j | s_i = 0) \\
&= \sum_{d_j \in \{0,1\}, D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0 | d_{-j} = D_{-j}, d_j) p(d_{-j} = D_{-j}, d_j | s_i = 0) \\
&= p(d_j = 1 | s_i = 0) \sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0 | d_{-j} = D_{-j}, d_j = 1) p(d_{-j} = D_{-j} | s_i = 0) \\
&+ p(d_j = 0 | s_i = 0)) \sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0 | d_{-j} = D_{-j}, d_j = 0) p(d_{-j} = D_{-j} | s_i = 0)
\end{aligned}$$

Similarly as before, we have

$$\sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0 | d_{-j} = D_{-j}, d_j = 0) p(d_{-j} = D_{-j}, d_j | s_i = 0) = r_y$$

and

$$\sum_{D_{-j} \in \{0,1\}^{m-1}} p(s_y = 0 | d_{-j} = D_{-j}, d_j = 1) p(d_{-j} = D_{-j}, d_j | s_i = 0) = r_y p_y$$

Altogether, this gives (3.4.5).

The proof that (3.4.6) holds is completely analogous to the above.

Finally, the claim that only one of the roots is non-negative follows immediately from Vieta's formulas.

□

Proof of 119. Analogously as in the proof of 118, the vector a' satisfies $a'_z = \frac{(1-q_i)+q_i p_z}{p_z}$. Similarly, by (3.4.8), we have $q_i = \frac{\rho p_i}{\rho p_i + (1-\rho)}$. This implies the statement of the Lemma. □

3.4.5.1 Peeling-off step

Since Algorithm 17 needs to maintain both the PMI matrix and second order moments, whenever we "peel off" a disease, we have to adjust accordingly both the PMI matrix and the moments.

Adjusting the PMI matrix:

To remove the contribution of W_j from PMI, we can simply subtract $f(W_j W_j^\top)$ from PMI where f is defined in Lemma 115. The full algorithm for this is specified as Algorithm 19.

Adjusting the pairwise probabilities:

Maintaining the first and second moments is done by Algorithm 20. The proof of correctness is based on observing that as far as the marginals are concerned, peeling off a diseases j is equivalent to calculating the marginal probabilities in the model, conditioned on the $d_j = 0$. We formalize it below:

Algorithm 19 Subtract disease j

Inputs: Empirical PMI matrix $\hat{\text{PMI}}$, estimated weights W_j for a disease j.
Outputs: Estimate of PMI matrix with disease j subtracted.

1. Let $L = 10\frac{\log(\nu/\epsilon)}{\log(1/\rho)}$ and $\tau = \frac{\rho}{2}(1-\exp(-\nu))$. Define the matrix M as

$$M = \sum_{l=1}^{L} (-1)^{l+1} \frac{1}{l} \left(\frac{\rho}{1-\rho}\right)^{l} (1-\exp(-lW_j))(1-\exp(-lW_j))^{\top}$$

2. Return the matrix $\phi_\tau(\hat{\text{PMI}} - M)$ where $\phi_\tau(x) = x$, if $x > \tau$, and $\phi_\tau = 0$, otherwise.

Algorithm 20 Update pairwise correlations

Inputs: Empirical PMI matrix $\hat{\text{PMI}}$, empirical estimates of $\hat{p}(\bar{s}_i, \bar{s}_k)$, $\hat{p}(\bar{s}_i)$, estimate of disease W_j
Outputs: New empirical estimates of $\tilde{p}(\bar{s}_i, \bar{s}_k)$, $\hat{p}(\bar{s}_i)$, with disease j removed.

For each pair of symptoms i, k, update the first and second moment by

$$\tilde{p}(\bar{s}_i, \bar{s}_k) = \frac{p(\bar{s}_i, \bar{s}_k)}{1-\rho+\rho(1-\exp(-W_{i,j}))(1-\exp(-W_{k,j}))}$$

$$\tilde{p}(\bar{s}_i) = \frac{p(\bar{s}_i)}{1-\rho+\rho(1-\exp(-W_{i,j}))}, \text{ and } \quad \tilde{p}(s_i, \bar{s}_k) = \tilde{p}(\bar{s}_i) - \tilde{p}(\bar{s}_i, \bar{s}_k)$$

Lemma 120. *Given the values of the pairwise marginal probabilities $p(\bar{s}_i, \bar{s}_k)$ for all pairs, Algorithm 20 calculates the marginal probabilities $\tilde{p}(i, k|d_j = 0)$.*

Proof. Consider first $p(\bar{s}_i, \bar{s}_k|d_j = 0)$. By (3.3.1), we have

$$\frac{p(\bar{s}_i, \bar{s}_k|d_k = 1)}{p(\bar{s}_i, \bar{s}_k|d_j = 0)} = (1-\exp(-W_{i,j})(1-\exp(-W_{k,j})$$

Since

$$p(\bar{s}_i, \bar{s}_k|d_j = 1)p(d_j = 1) + p(\bar{s}_i, \bar{s}_k|d_j = 0)p(d_j = 0) = p(\bar{s}_i, \bar{s}_k)$$

and $p(d_j = 1) = \rho$, we have

$$p(\bar{s}_i, \bar{s}_k|d_j = 1) = \frac{p(s_i, s_k)}{1-\rho+\rho(1-\exp(-W_{i,j}))(1-\exp(-W_{k,j}))}$$

which implies

$$\tilde{p}(\bar{s}_i, \bar{s}_k|d_j = 1) = \frac{p(s_i, s_k)}{1-\rho+\rho(1-\exp(-W_{i,j}))(1-\exp(-W_{k,j}))}$$

Completely analogously,

$$\tilde{p}(\bar{s}_i) = \frac{p(s_i)}{1-\rho+\rho(1-\exp(-W_{i,j}))}$$

The equality

$$p(s_i, \bar{s_k}|d_j = 0) = p(\bar{s_i}|d_j = 1) - p(\bar{s_i}, \bar{s_k}|d_j = 1)$$

implies

$$\tilde{p}(s_i, \bar{s_k}) = \tilde{p}(\bar{s_i}) - \tilde{p}(\bar{s_i}, \bar{s_k})$$

□

Combining these claims with Lemma 118 and 119, we get Theorem 20. This shows if we have the exact PMI matrix and pairwise correlations, the algorithm can recover the weight matrix exactly.

3.4.5.2 Error bounds for the components of Algorithm 17

In practice, we can only estimate the PMI matrix and the pairwise correlations with finite accuracy. To that end, we show that the algorithm is robust to noise.

Before we state the guarantees, we need to define an important parameter $p_{\min}$. Let D_t be the diseases peeled off up to stage t. Then $p_{\min}$ is defined as

$$p_{\min} = \min\left(\min_{i,k} p(s_i = 0, s_k = 0), \min_{D_t, i, k \notin D_t} p(s_i = 0|s_k = 1, d_{D_t} = 0)\right)$$

To parse this quantity, note that marginals that include conditioning on $d_{D_t} = 0$ are the same as the marginals in the network where we have peeled off the symptoms D_t. So, effectively this quantity is the minimum of the marginals $p(s_i = 0, s_j = 0)$ and $p(s_i = 0|s_j = 1)$ in either the original network, or after we have peeled off some of the symptoms. Now we can show that Algorithm 17 is robust to noise:

Theorem 21. *Suppose Assumptions A1 and A2 hold with greedy peeling depth T. Further, if all non-zero weights satisfy $W_{i,j} \geq \nu$ for some constant $\nu > 0$, Algorithm 17 recovers the parameters W up to error ϵ with high probability, given $O\left(\frac{1}{\epsilon^2}\left(\frac{\rho n}{p_{\min}}\right)^{2T}\right)$ samples. Moreover, the algorithm [10] runs in time $O\left(Tn^3\right)$.*

We have already shown that Algorithm 17 works with exact input. To make it robust, we need to make each step robust and bound how the errors propagate.

To that end, we first state a few lemmas that show how the error propagates in different steps of the algorithm, which we will combine these steps to prove the main theorem.

The lemmas needed for analyzing errors in the individual steps of the algorithm are the following:

[10] Excluding the time to construct the empirical PMI matrix

Lemma 121 (Errors in the recovery of the W parameters)**.** *If the pairwise probabilities $p(s_i, s_k), \forall i, k$ are accurate to within additive error ϵ, then the result of Algorithm 18 recovers the weights for W_j for all diseases with at least two anchors to within additive error $O(\epsilon/p_{\min})$.*

Lemma 122 (Errors when subtracting a disease)**.** *If the weight parameters W_j and the entries of the PMI matrix are accurate to within an additive error ϵ, after peeling off a diseases j via Algorithm 19, the entries of the PMI matrix are accurate to within additive error $O(\rho\epsilon)$. Consequently, if all the weight parameters have additive error ϵ, and a pair of symptoms appear in at most N diseases, after subtracting them all, the entries are accurate to $O(\rho N\epsilon)$. Furthermore, if $N\epsilon = o(1)$, the support of the entries of the PMI matrix are correct.*

Lemma 123 (Errors when adjusting probabilities when subtracting a disease)**.** *If the pairwise and singleton probabilities $p(s_i = 0, s_k = 0), \forall i, k, p(s_i = 0), \forall i$, as well as the parameters W_j for a disease g are accurate to within additive error ϵ, after peeling off a disease j by Algorithm 20, the new pairwise pairwise and singleton probabilities are accurate to within additive error $\epsilon + O(\rho\epsilon)$.*

We proceed to the proofs:

Proof of Lemma 121. Consider $W_{i,j}$ first. Algorithm 18 first recovers $1 - \exp(-W_{i,j})$ by solving a quadratic equation in q in Step 3. Rearranging this equation to write it in canonical form, we have that q is the solution to

$$r_2\rho q^2 + (r_2 - r_1)(1-\rho)q - r_1(1-\rho) = 0$$

where $r_2 = \frac{\hat{p}(s_i=0|s_k=0)}{\hat{p}(s_i=0|s_k=1)}$ and $r_1 = \frac{R_{-1}b}{R_{-1}a}$. By the usual quadratic formula,

$$q = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

where $B = (r_2 - r_1)(1-\rho), A = r_2\rho, C = r_1(1-\rho)$. Now, the following elementary inequalities for how the accuracy behaves under various arithmetic operations are very easy to show:

$$x^2 - 2\Delta \leq (x \pm \Delta)^2 \leq x^2 + 3\Delta, \text{ for } \Delta, x < 1 \tag{3.4.12}$$

$$xy - 2\Delta \leq (x \pm \Delta)(y \pm \Delta) \leq xy + 3\Delta, \text{ for } \Delta, x, y < 1 \tag{3.4.13}$$

$$\frac{1}{x} - \frac{\Delta}{x} \leq \frac{1}{x \pm \Delta} \leq \frac{1}{x} + (1 + e/2)\frac{\Delta}{x}, \text{ for } \Delta < x \tag{3.4.14}$$

$$\sqrt{x} - \sqrt{\Delta} \leq \sqrt{x \pm \Delta} \leq \sqrt{x} + \sqrt{\Delta}, \text{ for } \Delta < x \tag{3.4.15}$$

From these it is clear that under the assumptions of the Lemma, q will be accurate up to error $O(\epsilon/\gamma)$. Similarly as in the proof of Lemma 122, $\log(x)$ is $O(1)$-Lipschitz in the region $[\nu, +\infty]$ for any constant ν, so $W_{i,j}$ will be accurate up to error $O(\epsilon/\gamma)$.

The argument for $W_{z,j}, z \neq i$ is analogous: all the arithmetic operations in Step 5 are again subsumed by the above inequalities, so the a $O(\epsilon/\gamma)$ bound holds as well.

□

Proof of Lemma 122. We consider the i,k-th entry of the PMI-matrix. If $W_{i,j} = 0$ or $W_{k,j} = 0$, the peeling off procedure does not influence the entry and ϵ' will not change. By Lemma 115, the correct term to subtract is

$$\sum_{l=1}^{\infty} (-1)^{l+1} \frac{1}{l} \left(\frac{\rho}{1-\rho}\right)^l (1 - \exp(-lW_{i,j}))(1 - \exp(-lW_{k,j}))$$

Now we have estimates $W'_{i,j}, W'_{k,j}$ for $W_{i,j}$ and $W_{k,j}$, and $|W'_{i,j} - W_{i,j}| \leq \epsilon$, $|W'_{k,j} - W_{k,j}| \leq \epsilon$. The term we actually subtract (prior to thresholding at τ) is

$$\sum_{l=1}^{L} (-1)^{l+1} \frac{1}{l} \left(\frac{\rho}{1-\rho}\right)^l (1 - \exp(-lW'_{i,j}))(1 - \exp(-lW'_{k,j})),$$

so the error comes from truncating the Taylor series, and from the errors in the estimates W.

Consider the first error, let's call it δ_1. The function $\exp(-x)$ is 1-Lipschitz when $x > 0$, therefore $|\exp(-x) - \exp(-y)| \leq |x-y|$ for any x, y. Using $|(A+\Delta)(B+\Delta) - AB| \leq (A+B+|\Delta|)|\Delta|$, and the fact that $\exp(-lW_{i,j}), \exp(-lW_{k,j}) \in [0,1]$ we have

$$\begin{aligned}\delta_1 &\leq \sum_{l=1}^{L} \frac{1}{l} \left(\frac{\rho}{1-\rho}\right)^l \cdot 3l\epsilon \leq \sum_{l=1}^{\infty} \frac{1}{l} \left(\frac{\rho}{1-\rho}\right)^l \cdot 3l\epsilon \\ &\leq \sum_{l=1}^{\infty} 3\epsilon \left(\frac{\rho}{1-\rho}\right)^l = 3\epsilon \frac{\rho}{1-2\rho} \leq 4\rho\epsilon.\end{aligned}$$

Here the last inequality assumes $\rho \leq 1/8$ so that $1 - 2\rho \geq 3/4$.

Next, consider the second error, let's call it δ_2. We have

$$\begin{aligned}
\delta_2 &\leq \sum_{l=L+1}^{\infty} (-1)^{l+1} \frac{1}{l} \left(\frac{\rho}{1-\rho}\right)^l (1 - \exp(-lW_{i,j}))(1 - \exp(-lW_{k,j})) \\
&\leq \sum_{l=L+1}^{\infty} (-1)^{l+1} \frac{1}{l} \left(\frac{\rho}{1-\rho}\right)^l (1 - \exp(-\nu))^2 \\
&\leq \left(\frac{\rho}{1-\rho}\right)^{L+1} \frac{1-\rho}{1-2\rho} (1 - \exp(-\nu))^2 \\
&\leq \frac{4}{3} \left(\frac{\rho}{1-\rho}\right)^{L+1} (1 - \exp(-\nu))^2
\end{aligned}$$

By our choice of L, $\delta_2 = O(\rho\epsilon)$. The part of the lemma that concerns subtracting multiple diseases follows from induction on the number of peeled off diseases, and notice that an entry (i, k) does not get effected if $W_{i,j} = 0$ or $W_{k,j} = 0$.

Now, finally, note by the Taylor expansion (in Lemma 115), if a term in the PMI matrix is non-zero, it has magnitude at least $\frac{1}{10}\rho(1 - \exp(-\nu))^2$ for ρł/8, which implies that when $N\epsilon = o(1)$, the support will be correctly determined, which completes the proof of the lemma. □

Proof of Lemma 123. The proof follows a similar outline as the proof of Lemma 122. Consider, for instance $\tilde{p}(s_i, s_k)$: the quantities $\frac{1}{1-\rho+\rho(1-\exp(-W_{i,j}))(1-\exp(-W_{k,j}))}$ are accurate up to $O(\rho\epsilon)$ additive, and $p(s_i, s_k)$ is accurate up to ϵ, so $\tilde{p}(s_i, s_k)$ is accurate up to $\epsilon + O(\rho\epsilon)$.

The singleton marginals are handled the same way. □

Finally, we put together all these error bounds to get the final guarantee in Theorem 21:

Proof of Theorem 21. We will prove by induction on the the number of peeling off stages the following statement: at stage t, the support of the PMI matrix is correct, and the values $\hat{p}(\bar{s}_i, \bar{s}_k)$, $\hat{p}(\bar{s}_i)$, and the entries of the PMI matrix are accurate up to error $O(\epsilon\left(\frac{\rho N}{p_{\min}}\right)^t$.

Consider the base step first. By Lemma E.1 in (Arora et al., 2017b), if the number of samples is $\Omega(\frac{1}{p_{\min}} \log n \frac{1}{\epsilon^2})$, the empirical PMI matrix, as well as the pairwise marginals are entrywise accurate up to ϵ. Furthermore, Lemma 122 implies the thresholding step will correctly recover the supports of the PMI matrix.

Now, consider the inductive step. Towards that, suppose the statement is correct at time step t. Since the support of the PMI matrix is correct, by Lemma 117, the anchor symptoms will be correctly identified. But, given this, by Lemmas 118, 119 and 121, the weights of the diseases with at least two anchors will be recovered with accuracy $O\left(\epsilon\left(\frac{\rho N}{p_{\min}}\right)^{t+1}\right)$. Finally, by Lemma 123, the values $\hat{p}(\bar{s}_i, \bar{s}_k)$, $\hat{p}(\bar{s}_i)$ are accurate up to $O\left(\epsilon\left(\frac{\rho N}{p_{\min}}\right)^{t+1}\right)$ (since there are at most N diseases at stage t), and by Lemma 122 the PMI matrix values are accurate up to $O\left(\epsilon\left(\frac{\rho N}{p_{\min}}\right)^{t+1}\right)$ and the support is recovered correctly – thus finishing the proof of the theorem.

□

3.4.6 A generative model to understand the algorithm

The disease-symptom connections in QMR-DT interesting properties such as the ones exploited in the "peeling"algorithms of the current paper and also earlier by (Jernite et al., 2013). This section suggests a way to think about the structure of such networks using a generative model for the edges connecting related diseases and symptoms, such that it has these properties. By contrast, usual random graph models (such as the one assumed in (Arora et al., 2017b)) lack this structure. We hope this model may be useful in thinking about more applications for the algorithm.

The generative model is a random graph model for generating a bipartite graph. It assumes that the diseases as well as symptoms enter the graph in *epochs*. (Each epoch ends up being one of the stages in the 'peeling"during the algorithm.) Initially there is a set of diseases and symptoms that happen to be connected by a sparse random graph, where the sparsity is such that all diseases have anchor symptoms during the epoch. In each epoch a new set of diseases enter the graph, as do a new set of symptoms. The diseases take up some symptoms from the existing set of symptoms, and some from the new set of symptoms. Since the newly arrived diseases took on old symptoms, now some of those symptoms may stop being anchor symptoms if they now occur in two diseases. But the newly arrived diseases have anchor symptoms among the new pool of symptoms that also arrived during the epoch. (Thus intuitively speaking, our peeling algorithm will peel off epochs in reverse order that they arrived in.)

Now we make this model precise. Concretely, at epoch T, we will have m_T diseases and n_T symptoms, and each symptom will be included in the support of one of the diseases with probability $p_{T,T}$.

At each previous epoch t, we will increase the set of diseases by adding m_t new diseases and n_t symptoms to the already existing ones; each of the $n_i, i \in [t, T]$ symptoms will be included in the support of the new m_t diseases with probability $p_{t,i}$.

We will call a symptom an *anchor at epoch t*, if the symptom is an anchor symptom when only considering the diseases in epochs $[t, T]$. Similarly, a disease *has an anchor at epoch t*, if it has an anchor at epoch t.

For notational convenience, let's denote $m = \sum_{t=1}^{T} m_t$; furthermore, let's denote by M_t the set of diseases added at epoch t, and by N_t the set of symptoms added at epoch t. We will say an event happens *with high probability* if it happens with probability $1 - \exp(-\Omega(\log^2 m))$.

The range of the parameters for the generative model will be as follows:

(G1) $p_{t,t} n_t = \Theta(\log^2 m)$ and $n_t = \omega(m_t \log^2 m)$

(G2) $p_{t,t} \leq \frac{1}{2m_t}$ and $m_t = \omega(\log^2 m)$

(G3) $\sum_{i=1}^{T} 1/m_t \leq 1$, and $p_{t,i} m_t = \Omega(\log^2 m)$

Graphically, the generative model is represented in Figure 3.2. Ochre denotes the symptoms that are anchors at that epoch, while maroon denotes symptoms that are *not* anchors up to that epoch.

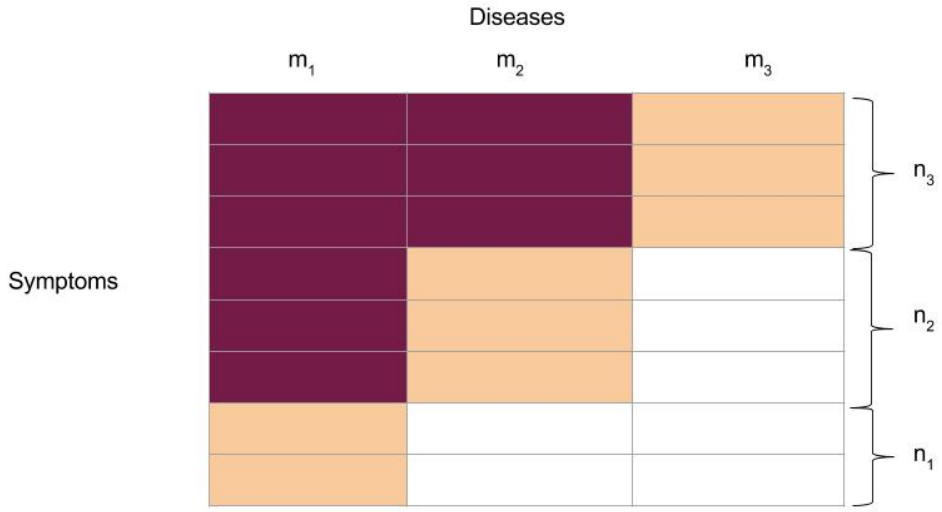


Figure 3.2: Pictorial representation of generative model. Ochre denotes symptoms that are anchors at that epoch, maroon symptoms that are *not* anchors up to that epoch.

We will show the following properties of our generative model:

Theorem 22. *Under the setup in this section, we have:*

(i) With high probability, the diseases M_t have anchors at epoch t.

(ii) With high probability, the symptoms N_t are not anchors at epoch $t' < t$.

(iii) Assumption (A1) is satisfied.

We note that (i) and (iii) are the only conditions relevant to the algorithm's correctness, but (ii) is empirically true as well, so we incorporate it in our generative model. We further note that by setting the parameters accordingly, we can also ensure that each disease has roughly the same number of symptoms in addition to (i)-(iii). Namely, we claim that there is a setting of the parameters, such that all diseases have roughly same number of symptoms: $\forall t : \sum_{i=1}^{t} n_i p_{t,i} = \Theta\left(\log^2 m\right)$. One particular setting is:

- $m_1 = \log^3 m$ and $m_t = 2\log^5 m \cdot m_{t-1}$, for some other constant C. (Accordingly, the number of epochs is $T = O(\log m / \log\log m)$)
- $n_t = m_t \log^3 m$
- $p_{t,t} = \frac{1}{m_t \log m}$ and $p_{t,i} = \frac{1}{2m_t}$, for $i < t$ and $t > 1$.

For these settings, $\sum_{i=1}^{t} n_i p_{t,i} = \sum_{i=1}^{t-1} n_i p_{t,i} + p_{t,t} n_t = \log^2 m \left(\sum_{i=1}^{t-1} m_i p_{t,i}\right) + \log^2 m$. But, note that $m_i p_{t,i} = (2\log^5 m)^{i-t}$, so $\sum_{i=1}^{t-1} m_i p_{t,i} \le 1/\log^5 m$, which implies that $\sum_{i=1}^{t} n_i p_{t,i} = \Theta(\log^2 m)$ for all t.

3.4.7 Experimental results

Finally, we turn to practical aspects of our approach. The algorithm is quite fast: a vanilla implementation in Matlab runs in under an hour on a dual, eight-core 2.8GHz Intel Xeon E5 2680 v2 processor machine on the QMR-DT dataset, which has 4075 symptoms and 570 diseases. In contrast, (Jernite et al., 2013) do not report run times for their algorithm on QMR-DT – only on a much smaller dataset (Singliar and Hauskrecht, 2006) with 8 latent and 64 observable variables.

Additionally, there are many operations in the algorithm (e.g. finding duplicate anchor symptoms for a particular disease) that in principle could be substantially sped up using hashing tricks. With respect to the peeling depth: we find 8 layers of peeling suffice to recover all but 2 diseases using our algorithm. This problematic disease pair —they share all but 1 symptom—was already identified in (Halpern and Sontag, 2013)[11]

We can run our algorithm with real samples, but sample complexity is a problem due to numerical issues. Using infinite samples (i.e., the correct PMI matrix), the algorithm should identify 229 diseases in the first iteration, 132 in the second, 83 in the third. When we estimate the PMI matrix with 100 million samples, 90% of the diseases can be correctly identified in the first iteration with almost no false positives, and 70% in the second one again with few false positives. (Thus 300 of the 570 diseases were correctly identified with samples.) However, in the third round, the precision/recall tradeoff significantly nosedives, and significant experimentation did not improve this.

[11] In principle, if we had an alternate procedure for peeling off diseases using only a single anchor even fewer rounds would suffice – on QMR-DT, 4 would be sufficient to recover all the diseases.

Chapter 4

Provable guarantees for inference in undirected, fully observable graphical models

In this chapter, we will present new results on provable guarantees for inference in undirected graphical models. More precisely, we will give guarantees for variational methods in the context of calculating partition functions. Our algorithms will be based on convex relaxations of the optimization problem that arises in variational approaches.

First, in Section 4.1 we will survey traditional approaches for calculating partition functions based on variational methods – both those with provable guarantees, and those without. In Section 4.3.1 we will describe our new relax-and-round approach to these methods using convex relaxations.

In Section 4.3 we will derive new algorithms for coarse multiplicative approximations of the partition function of Ising models, without any assumptions on the parameters of the model. These can be viewed as partition function analogues of Goemans-Williamson type results in the context of optimization. This section is based on results in (Li and Risteski, 2016).

In Section 4.4 we will improve the approximation guarantees of Section 4.3 when the graph of the Ising model has more structure. More precisely, we will show that for dense and spectrally well-behaved graphs, we can design much better algorithms. This section is based on results in (Risteski, 2016).

4.1 Overview of variational methods for calculating partition functions

We will briefly survey variational approaches to calculating partition functions. For simplicity, we will focus on Ising models, though the results are fully general. Recall, an Ising model is a distribution $p : \{-1,1\}^n \to [0,1]$ that has the form $p(x) \propto \exp\left(\sum_{i\sim j} J_{i,j} x_i x_j\right)$ and its partition function is $Z = \sum_{x\in\{-1,1\}^n} \left(\sum_{i\sim j} J_{i,j} x_i x_j\right)$.

The main idea all the algorithms use is the following simple lemma, which characterizes Z as the solution of an optimization problem. The idea is essentially the same as Lemma 3. In machine learning, this approach was rediscovered in efforts to understand belief propagation (Wainwright and Jordan, 2008; Yedidia et al., 2003). Concretely, the following claim holds:

Lemma 124 (Variational characterization of $\log Z$)**.** *For any distribution $q : \{-1,1\}^n \to [0,1]$,*

$$\sum_{i\sim j} J_{i,j}\mathbb{E}_q\left[x_i x_j\right] + H(q) \leq \log Z$$

with equality at $q = p$.

Proof. For any distribution $q : \{-1,1\}^n \to [0,1]$, we can write the KL divergence between q and p as

$$KL(q\|p) = \mathbb{E}_q\left[\log q(x)\right] - \mathbb{E}_q\left[\log p(x)\right] = -H(q) - \sum_{i\sim j} J_{i,j}\mathbb{E}_q\left[x_i x_j\right] + \log Z$$

Since the KL divergence is always non-negative, $-H(q) - \sum_{i\sim j} J_{i,j}\mathbb{E}_q\left[x_i x_j\right] + \log Z \geq 0$. Hence, $\log Z \geq H(q) + \sum_{i\sim j} J_{i,j}\mathbb{E}_q\left[x_i x_j\right]$ which proves the first claim of the lemma. However, equality is achieved whenever the KL divergence is 0, which happens when $q = p$. This finishes the second part of the lemma. □

An immediate consequence of the above is the following:

Corollary 125. $\log Z = \max_{q\sim\{-1,1\}^n}\left\{\sum_{i\sim j} J_{i,j}\mathbb{E}_q\left[x_i x_j\right] + H(q)\right\}$.

Of course, the above optimization problem is intractable, so the problem needs to be relaxed in some way. There are a few ways this has been done hitherto, which we summarize in the subsequent sections.

4.1.1 Constraining the distribution to optimize over

The first approach towards solving the optimization in Corollary 125 is analogous to the one used in variational Bayes in Section 2.1 – namely to optimize over a constrained class of distributions $\mathcal{Q}$. A common choice is *product distributions* – this is usually called a *mean-field* approximation. To make this concrete, note that the mean-field

distribution can be parametrized by the *single-variable* marginals q_i. So, the optimization to solve becomes:

$$\max_{q_i \sim \{-1,1\}, i \in [n]} \left\{ \sum_{i \sim j} J_{i,j} \mathbb{E}_{q_i}[x_i] \mathbb{E}_{q_j}[x_j] + \sum_{i=1}^{n} H(q_i) \right\} \tag{4.1.1}$$

Note that since each of the q_i distributions have $\{-1, 1\}$ as their domain, it is trivial to parametrize them, by say, $q_i(x_i = -1)$, so the above optimization problem involves $O(n)$ variables only, and the objective function can be efficiently evaluated. Though (4.1.1) is in general non-convex, practitioners frequently run iterative algorithms like gradient descent. Provable guarantees however are extremely rare: for instance, in the Curie-Weiss Ising model (in which the graph G is complete and $J_{i,j} = J \forall i, j$), for sufficiently small J the objective (4.1.1) is convex, and one can (coarsely) bound the quality of the approximation. (Friedli and Velenik, 2017)

4.1.2 Polytope-based approximations

Another family of approaches proceeds by reformulating the objective in Corollary 125 in terms of the low-order marginals of the distribution, and subsequently relaxing the polytope of marginals in some manner. To make this concrete, note that we can rewrite the objective in Corollary 125 as

$$\log Z = \max_{\tilde{q}_S(x_S), |S| \leq 2 \in \mathcal{M}^2} \left\{ \sum_{i,j \in E(G)} J_{i,j} \mathbb{E}_{\tilde{q}_{\{i,j\}}}[x_i x_j] + \max_{q \sim \{-1,1\}^n : q_S(x_S) = \tilde{q}(x_S), |S| \leq 2} H(q) \right\} \tag{4.1.2}$$

where $\mathcal{M}^2$ denotes the polytope of marginals over subsets of size 2.

In words, we can *first* optimize of the values of the pairwise marginals – which determines the value of the quadratic portion of the objective and subsequently maximize the entropy of q subject to matching these marginals. While this may seem like a trivial rewrite, this separate treatment of the entropy is very beneficial. Note that at this point, there are two sources of difficulty in solving (4.1.2): optimization over the polytope of pairwise marginals is still intractable; so is evaluating the quantity $H_{\max}(\tilde{q}^2)$.

Moment-based relaxations : The first difficulty is addressed by using *moment-based convex relaxations* of the polytope of pairwise marginals. The idea behind these families of relaxations is to introduce variables for the marginals, and impose certain constraints that would be satisfied by *valid* pairwise marginals. This is a technique which has received a lot of recent study in theoretical computer science as well – where two families of relaxations, the *Sherali-Adams convex programming hierarchy* and the *Lasserre/Parrillo/Sum-of-Squares hierarchy* have been extensively studied in the context of combinatorial optimization. We will survey these methods detail in Section 4.2.1, but for the sake of introducing approximations of the entropy (which address the second difficulty), we will formally define the polytope corresponding to the basic Sherali-Adams relaxation of order 2, which we denote as $SA(2)$.

The polytope $SA(2)$ is described by variables $q_S(x_S), x_S \in \{-1,1\}^{|S|}, |S| \leq 2$ corresponding to local distributions over subsets $S \subseteq [n]$ of size at most 2. The following natural constraints are imposed:

- *Non-negativity*: $\forall S \subseteq [n], |S| \leq 2, x_S \in \{-1,1\}^{|S|}, q_S(x_S) \geq 0$.
- *Consistency*: $\forall i \neq j \in [n], \forall x_i \in \{-1,1\}, \sum_{x_j \in \{-1,1\}} q_{i,j}(x_i, x_j) = q_i(x_i)$
- *Marginalization*: $\forall i \in [n], \sum_{x_i \in \{-1,1\}} q_i(x_i) = 1$

It is clear that these constraints would be satisfied by any *valid* marginals q_S of a distribution over the hypercube, but the converse need not be true – the variables q_S may not correspond to marginals of any distribution. In other words, $\mathcal{M}^2 \subseteq SA(2)$. However, since the number of variables in $SA(2)$ is $O(n^2)$, and the constraints we are imposing are linear, we can in polynomial time optimize convex functions over the polytope of such variables.

Entropy approximations : The second difficulty is that the above relaxations merely contain variables for marginals over sets of size 2. Thus it is unclear how to evaluate the quantity $H_{\max}(\tilde{q}^2)$. In fact, even if the values $q_S(x_S)$ were valid marginals (which they need not be), this corresponds to a *maximum entropy* evaluation problem, subject to second order marginal constraints, which is well known to be NP-hard to even approximate (**?**) A popular approach for tackling this issue is to use the *Bethe approximation* of the entropy. For a distribution q, it is defined as

$$H_{\text{Bethe}}(q) = \sum_{i \sim j} H(q_{i,j}) - \sum_i (d_i - 1) H(q_i) \tag{4.1.3}$$

where d_i is the degree of vertex i. Note the above quantity depends only on the pairwise marginals – so is well defined and can be efficiently evaluated even for variables $\tilde{q}^2$ in SA(2). Thus, when clear from the context, we will abuse notation and refer to $H_{\text{Bethe}}(\tilde{q}^2)$. Thus, the relaxation to 4.1.2 we are solving

$$\max_{\tilde{q}^2 \in SA(2)} \sum_{i,j \in E(G)} J_{i,j} \mathbb{E}_{\tilde{q}_{\{i,j\}}}[x_i x_j] + H_{\text{Bethe}}(\tilde{q}^2) \tag{4.1.4}$$

The quantity (4.1.3) may seem strange at first sight. The rationale behind it is that if the graph G is a tree, the entropy of the Ising distribution p corresponding to it is exactly $H_{\text{Bethe}}(p)$. Thus, the hope is that for graphs that are "tree-like", a notion which we will make formal shortly, this approximation is not too far from being accurate. Two important issues arise:

- $H_{\text{Bethe}}(p)$ is in general neither an upper nor a lower bound on $H(p)$. As a consequence, $H_{\text{Bethe}}(\tilde{q}^2)$ is neither an upper nor a lower bound of $H_{\max}(\tilde{q}^2)$. As a result, (4.1.4) is not a "relaxation" in the standard sense of theoretical computer science.

- The resulting objective function is not necessarily concave – so even making claims as to whether a global optimum of the optimization is reached is difficult in general. (In fact, the only known case when the objective is concave are graphs with a single cycle (Weiss, 2000).)

Formally, the provable guarantees on this approximation proceed via analyzing a particular iterative heuristic to solve (4.1.4) called *Belief Propagation*. A technical description of this procedure is beyond the scope of this thesis, but surveys on the variational viewpoint of belief propagation can be found in (Yedidia et al., 2003). We will however mention that the class of graphs G for which provable guarantees are known satisfy the following conditions:

- Locally tree-like: G has constant degree, and has large girth, i.e. no cycles of length $O(\log n)$.
- Correlation decay: far away nodes in the graph don't influence much the marginals at a particular node. More precisely, the marginal distributions at a node, conditioned on any two different values of the neighbors at distance t, differ by at most $O(e^{-t})$.

4.1.3 Advanced methods

Variational methods are still a very active research area, and as a consequence there are many variations on the above approaches, but most of which come with few guarantees. We will briefly survey on a high level the main "tweaks and tricks", with a particular accent on the classes of graphs for which provable guarantees are known.

Higher-order analogues of the Bethe entropy : If the moment-relaxation polytope includes variables for higher-order marginals – say, for sets of size k, similar "consistency" constraints as those in SA(2) can be imposed (see Section 4.2.1), and it is clear that this would be a tighter relaxation of the moment polytope. However, it is unclear how one can leverage this information to improve the quality of the Beth approximation. Heuristic arguments, based on the inclusion-exclusion principle and attempting "not to overcount" the local entropies lead to a variety of higher-order analogues, the most famous of which are the Kikuchi family of approximations. (Yedidia et al., 2003)

Convexifications of the Bethe entropy : One of the undesirable aspects of the Bethe approximation is that H_{Bethe} is not necessarily convex. There have been some efforts to provide *convexified* variants of this approximation: (Wainwright et al.; 2005) suggested a concave approximation based on convex combinations of tree entropies in the spanning trees of the graph; this approximation has the added property that it is an *upper bound* on $H_{\max}$. (Meshi et al., 2009) explored ways to find the "nearest" convexification of the Bethe approximation under various notions of closeness. Neither of these papers provide guarantees for the quality of the approximation on any class of graphs. (Additionally, empirical results suggest these convexifications frequently give much worse approximations to the partition function/marginals of the graph than the Bethe approximation.)

Loop-correcting the Bethe entropy : (Chertkov and Chernyak, 2006) provided a *loop-series* interpretation of the Bethe approximation, which allows expressing *exactly* the partition function as a function of any fixed point of the belief propagation iterations. Unfortunately, this expression requires iterating over all (simple) cycles in the graph G, hence is provably efficient only for graphs with a small (polynomial) number of simple cycles. Techniques of this kind were also used in the context of graphs with large girth, and counting independent sets (Chandrasekaran et al., 2011).

4.2 Our approach: rounding and entropy approximations

Our approach will follow the *moment-based* relaxation paradigm as far as relaxing the polytope of marginals is concerned; where we will diverge is how we handle the entropy portion of the objective. We will construct entropy approximations that remedy both issues with the Bethe entropy; namely, they will upper bound $H_{\max}$ and they will be concave. Subsequently, we will *round* the solution obtained to an actual distribution, in a manner that doesn't lose too much in terms of the value of the objective function. In this sense, our approach is much more akin to usual relax-and-round approach of linear/semidefinite relaxations to combinatorial optimization problems. Concretely, the simple proposition we use is the following:

Proposition 126. *Let $\mathcal{M}^2 \subseteq \mathcal{M}'$, and let the functionals $\tilde{H}, \tilde{E} : \mathcal{M}' \to \mathbb{R}$ satisfy $H(\tilde{q}) \leq \tilde{H}(\tilde{q}), \mathbb{E}_{\tilde{q}}\left[x_i x_j\right] = \tilde{\mathbb{E}}_{\tilde{q}}\left[x_i x_j\right]$ for $\tilde{q} \in \mathcal{M}$, then*

$$\log Z \leq \max_{\tilde{q} \in \mathcal{M}'} \left\{ \sum_{i \sim j} J_{i,j} \tilde{\mathbb{E}}_{\tilde{q}}\left[x_i x_j\right] + \tilde{H}(\tilde{q}) \right\}$$

Subsequently, we will *round* the pseudo-distributions to actual distributions, in a manner that doesn't lose too much in terms of the value of the objective function. What this means is that we will provide an *actual* distribution q, for which

$$\sum_{i \sim j} J_{i,j} \mathbb{E}_q\left[x_i x_j\right] + H(q)$$

is comparable to

$$\sum_{i \sim j} J_{i,j} \tilde{\mathbb{E}}_{\tilde{q}}\left[x_i x_j\right] + \tilde{H}(\tilde{q})$$

4.2.1 Convex programming hierarchies

We will introduce the general family of polytopes we will use for approximating the marginals polytope. we provide only a (very) brief overview for completeness sake, and will not attempt to give a thorough survey of this rich area of study. For more details, the reader can consult (Barak et al., 2011; 2014; Laurent, 2009).

The k-level *Sherali-Adams* hierarchy polytope (henceforth SA(k)) is defined by variables $\tilde{q}_S(x_S), x_S \in \{-1,1\}^{|S|}$ specifying local distributions over all subsets $S \subseteq [n]$, $|S| \leq k$. The distributions $\tilde{q}_S : \{-1,1\}^{|S|} \to [0,1]$ and $\tilde{q}_T : \{-1,1\}^{|T|} \to [0,1]$, for any S, T s.t. $|S \cup T| \leq k$ must be "consistent" on $S \cap T$. More precisely, it's the case that

$$\Pr_{x_S \sim \tilde{q}_S}[x_{S\cap T} = \alpha] = \Pr_{x_T \sim \tilde{q}_T}[x_{S\cap T} = \alpha], \forall S, T \subseteq [n], |S \cup T| \leq k$$

The fact that these constraints can be written as a linear program was sketched in Section 4.1.2 for $k = 2$; for more details, the reader can consult (Barak et al., 2011).

We can also define a *conditioning* operation thanks to the existence of these local distributions. More precisely, for a vertex v, *conditioning* on v involves sampling v according to the local distribution $\tilde{q}_{\{v\}}$. This operation specifies a solution to the $k-1$-st level SA hierarchies: just define $\tilde{q}_S(x_S) = \tilde{q}_{S\cup\{v\}}(x_{S\cup v})$.

Moreover, a natural *pseudo-expectation* functional $\tilde{\mathbb{E}}_{\tilde{q}}[\cdot]$ can be defined for any polynomial of degree at most k, by defininining for monomials $\pi_{i\in I} x_i, I \subseteq [n]$,

$$\tilde{E}_{\tilde{q}}[\pi_{i\in I} x_i] = \Pr_{\tilde{q}}[\pi_{i\in I} x_i = 1] \tag{4.2.1}$$

and extending it to polynomials by linearity.

The polytope corresponding to the k-th level *Lasserre* or Sum-of-Squares hierarchy (henceforth LA(k)) is a semidefinite program s.t. there are vectors $v_{S,\alpha}$ for each subset S and possible assignment of values α to it satisfying $\langle v_{S,\alpha}, v_{T,\beta}\rangle = \Pr_{\tilde{q}_{S\cup T}}(x_S = \alpha, x_T = \beta)$, if $|S \cup T| \leq k$. The additional power that is gained by this is that the natural *pseudo-expectation* functional $\tilde{\mathbb{E}}_{\tilde{q}}[\cdot]$ satisfies

$$\tilde{E}_{\tilde{q}}[(r(x))^2] \geq 0$$

for any polynomial r of degree at most $k/2$.

4.3 Worst-case guarantees using approximate maximum entropy principles

To illustrate our approach, we will first tackle the task of giving *worst-case* guarantees for coarse approximations of the partition function of Ising models, with no assumptions on the structure of the potentials. More precisely, we tackle the following basic research question:

What is the best approximation guarantee on the partition function in the worst case (with no additional assumptions on the potentials)?

In the low-temperature limit, i.e. when $|J_{i,j}| \to \infty$, $\log Z \to \max_{x\in\{-1,1\}^n} \sum_{i\sim j} J_{i,j} J_{i,j} x_i x_j$ - i.e. the question reduces to purely to optimization. In this regime, this question has a very satisfying answer: in the worst case, one can get

$O(\log n)$ factor multiplicative factor approximation of the log of the partition function, and unless P = NP, one cannot get better than constant factor approximations of it (Charikar and Wirth, 2004; Alon and Naor, 2006; Alon et al., 2006)

In the finite-temperature version, it is known that it is NP-hard to achieve a $(1+\epsilon)$-factor approximation to the partition function (i.e. construct a FPRAS) (Sly and Sun, 2012), but nothing is known about coarser approximations. We prove in this section, informally, that one can get comparable multiplicative guarantees on the *log-partition function* in the finite temperature case as well. The methods are generic, and likely to apply to many other exponential families, where algorithms based on linear/semidefinite programming relaxations are known to give good guarantees in the optimization regime.

We will prove the following results:

Theorem 127 (Ferromagnetic Ising, informal)**.** *There is a convex optimization problem over LA(2) that calculates up to multiplicative approximation factor 50 the value of* $\log Z$ *where* Z *is the partition function of the Ising model* $p(x) \propto \exp(\sum_{i,j} J_{i,j} x_i x_j)$ *for* $J_{i,j} > 0$.

Theorem 128 (Ising model, informal)**.** *There is a convex optimization problem over LA(2) that calculates up to multiplicative approximation factor* $O(\log n)$ *the value of* $\log Z$ *where* Z *is the partition function of the exponential distribution* $p(x) \propto \exp(\sum_{i,j} J_{i,j} x_i x_j)$.

Theorem 129 (Ising model, informal)**.** *There is a convex optimization problem over LA(2) that calculates up to multiplicative approximation factor* $O(\log \chi(G))$ *the value of* $\log Z$ *where* Z *is the partition function of the exponential distribution* $p(x) \propto \exp(\sum_{i,j \in E(G)} J_{i,j} x_i x_j)$ *and* G *has chromatic number* $\chi(G)$.

Note Theorem 129 is strictly more general than Theorem 128, however the proof of Theorem 128 uses less heavy machinery and is illuminating enough that we feel merits being presented as a separate result.

In all of the above theorems, the entropy functional $\tilde{H}$ will be *trivial*: we will set $\tilde{H}(\tilde{q}) = n$. This is trivially an upper bound on $H_{\max}$, as needed by Proposition 126. The difficulty will be proving that the rounding can produce a distribution with nearly-maximum entropy. The idea is similar for all three results above, but the intuition is most readily illustrated in the ferromagnetic case, so we proceed to that one first.

4.3.1 Ferromagnetic Ising models

Denoting by $\mathcal{G} = \min_{t \in [-1,1]} \left\{ \frac{2}{\pi} \arcsin(t)/t \right\} \approx 0.64$, we will prove the following natural approximate maximum entropy principle:

Theorem 130 (Ferromagnetic, approximate entropy principle, (Li and Risteski, 2016))**.** *For any positive-semidefinite matrix* Σ *with* $\Sigma_{i,i} = 1, \forall i$*, there is an efficiently sampleable distribution* $q : \{-1,1\}^n \to \mathbb{R}$*, which can be sampled as*

sign(g)*, where* $g \sim \mathcal{N}(0, \Sigma + \beta I)$*, and satisfies* $\frac{\mathcal{G}}{1+\beta}\Sigma_{i,j} \leq E_q[x_i x_j] \leq \frac{1}{1+\beta}\Sigma_{i,j}$ *and has entropy* $H(q) \geq \frac{n}{25}\frac{(3^{1/4}\sqrt{\beta}-1)^2}{\sqrt{3\beta}}$*, where* $\beta \geq \frac{1}{3^{1/2}}$.

In order to parse the theorem, if we think of the covariance matrix as the moments of a distribution over $\{-1, 1\}$ for which the singleton marginals are 0, there is an *efficiently sampleable distribution* which preserves these moments up to a constant multiplicative factor, and has linear entropy (i.e. at most a constant factor away from the maximum possible entropy.) Of course, the theorem is true for any covariance matrix – i.e. the entries need not be actual moments.

This is an *approximate maximum entropy* principle, in the sense that the maximum entropy distribution matching certain second order moments is an Ising model. This is a well known fact (Jaynes, 1957) and is one of the justifications for the widespread use of Ising models in machine learning (and more generally, Markov Random Fields.) However determining the potentials of this Ising model, even approximately is impossible unless RP = NP. The above theorem is a "bi-criteria" approximation of this principle: i.e. it only approximately matches both the moments and the entropy of the distribution, but is efficient.

Before proving this theorem, we will see how it can be used to design an approximation algorithm for the partition function of a ferromagnetic Ising model. Recall the celebrated First Griffiths inequality due to Griffiths (Griffiths, 1967) which states that for ferromagnetic Ising models, $\mathbb{E}_p[x_i x_j] \geq 0, \forall i, j$.

Using this inequality, we will look at the following natural relaxation:

$$\max_{\tilde{q} \in \mathrm{LA}(2); \tilde{\mathbb{E}}_{\tilde{q}}[x_i x_j] \geq 0, \forall i,j} \left\{ \sum_{i \sim j} J_{i,j} \tilde{\mathbb{E}}_{\tilde{q}}[x_i x_j] + n \right\} \tag{4.3.1}$$

Theorem 131 ((Li and Risteski, 2016))**.** *The relaxation (4.3.1) provides a factor 50 approximation of* $\log Z$.

Proof. Due to Griffiths' inequality and Proposition 126, (4.3.1) is an upper bound of $\log Z$. We will provide a rounding of (4.3.1), as described in Section . We will use the distribution q from Lemma 130: the sign of a Gaussian with covariance matrix $\Sigma + \beta I$, for a β which we will specify. By Theorem 130, we then have $H(q) \geq \frac{n}{25}\frac{(3^{1/4}\sqrt{\beta}-1)^2}{\sqrt{3\beta}}$ whenever $\beta \geq \frac{1}{3^{1/2}}$ and $E_q[x_i x_j] \geq \dfrac{\mathcal{G}}{1+\beta}\tilde{\mathbb{E}}_{\tilde{q}}[x_i x_j]$

By setting $\beta = 21.8202$, we get $\frac{n}{25}\frac{(3^{1/4}\sqrt{\beta}-1)^2}{\sqrt{3\beta}} \geq 0.02$ and $\frac{\mathcal{G}}{1+\beta} \geq 0.02$, which implies that

$$\sum_{i,j} J_{i,j}\mathbb{E}_q[x_i x_j] + H(q) \geq 0.02\left(\sum_{i,j} J_{i,j}\tilde{\mathbb{E}}_{\tilde{q}}[x_i x_j] + n\right)$$

which implies the claim we want. □

With this in mind, we will prove Theorem 130. We will do this in two parts – by first lower bounding the entropy of $\tilde{\mu}$, and then by bounding the moments of $\tilde{\mu}$.

Theorem 132 ((Li and Risteski, 2016)). *The entropy of the distribution $\tilde{\mu}$ satisfies $H(\tilde{\mu}) \geq \frac{n}{25}\frac{(3^{1/4}\sqrt{\beta}-1)^2}{\sqrt{3\beta}}$ when $\beta \geq \frac{1}{3^{1/2}}$.*

Proof. A sample g from $\mathcal{N}(0, \tilde{\Sigma})$ can be produced by sampling $g_1 \sim \mathcal{N}(0, \Sigma)$, $g_2 \sim \mathcal{N}(0, \beta I)$ and setting $g = g_1 + g_2$. The sum of two multivariate normals is again a multivariate normal. Furthermore, the mean of g is 0, and since g_1, g_2 are independent, the covariance of g is $\Sigma + \beta I = \tilde{\Sigma}$.

Let's denote the random variable $\mathbb{Y} = \text{sign}(g_1 + g_2)$ which is distributed according to $\tilde{\mu}$. We wish to lower bound the entropy of $\mathbb{Y}$. Toward that goal, denote the random variable $\mathbb{S} := \{i \in [n] : |(g_1)_i| \leq cD\}$ for c, D to be chosen. Then, we have: for $\gamma = \frac{c-1}{c}$,

$$H(\mathbb{Y}) \geq H(\mathbb{Y}|\mathbb{S}) = \sum_{S \subseteq [n]} \Pr[\mathbb{S} = S]H(\mathbb{Y}|\mathbb{S} = S) \geq \sum_{S \subseteq [n], |S| \geq \gamma n} \Pr[\mathbb{S} = S]H(\mathbb{Y}|\mathbb{S} = S)$$

where the first inequality follows since conditioning doesn't decrease entropy, and the latter by the non-negativity of entropy. Continue the calculation we can get:

$$\begin{aligned}\sum_{S \subseteq [n], |S| \geq \gamma n} \Pr[\mathbb{S} = S]H(\mathbb{Y}|\mathbb{S} = S) &\geq \sum_{S \subseteq [n], |S| \geq \gamma n} \Pr[\mathbb{S} = S] \min_{S \subseteq [n], |S| \geq \gamma n} H(\mathbb{Y}|\mathbb{S} = S) \\ &= \Pr\left[|\mathbb{S}| \geq \gamma n\right] \min_{S \subseteq [n], |S| \geq \gamma n} H(\mathbb{Y}|\mathbb{S} = S)\end{aligned}$$

We will lower bound $\Pr[|\mathbb{S}| \geq \gamma n]$ first. Notice that $\mathbb{E}[\sum_{i=1}^n (g_1)_i^2] = n$, therefore by Markov's inequality, $\Pr\left[\sum_{i=1}^n (g_1)_i^2 \geq Dn\right] \leq \frac{1}{D}$. On the other hand, if $\sum_{i=1}^n (g_1)_i^2 \leq Dn$, then $|\{i : (g_1)_i^2 \geq cD\}| \leq \frac{n}{c}$, which means that $|\{i : (g_1)_i^2 \leq cD\}| \geq n - \frac{n}{c} = \frac{(c-1)n}{c} = \gamma n$. Putting things together, this means $\Pr\left[|\mathbb{S}| \geq \gamma n\right] \geq 1 - \frac{1}{D}$.

It remains to lower bound $\min_{S \subseteq [n], |S| \geq \gamma n} H(\mathbb{Y}|\mathbb{S} = S)$. For every $S \subseteq [n], |S| \geq \gamma n$, denote by $\mathbb{Y}_S$ the coordinates of $\mathbb{Y}$ restricted to S, we get

$$H(\mathbb{Y}|\mathbb{S} = S) \geq H(\mathbb{Y}_S|\mathbb{S} = S) \geq H_\infty(\mathbb{Y}_S|\mathbb{S} = S) = -\log(\max_{y_S} \Pr[\mathbb{Y}_S = y_S|\mathbb{S} = S])$$

(where H_∞ is the min-entropy) so we only need to bound $\max_{y_S} \Pr[\mathbb{Y}_S = y_S|\mathbb{S} = S]$

We will now, for any y_S, upper bound $\Pr[\mathbb{Y}_S = y_S|\mathbb{S} = S]$. Recall that the event $\mathbb{S} = S$ implies that $\forall i \in S$, $|(g_1)_i| \leq cD$. Since g_2 is independent of g_1, we know that for every fixed $g \in \mathbb{R}^n$:

$$\Pr[\mathbb{Y}_S = y_S|\mathbb{S} = S, g_1 = g] = \Pi_{i \in S} \Pr[\text{sign}([g]_i + [g_2]_i) = y_i]$$

For a fixed $i \in [S]$, consider the term $\Pr[\text{sign}([g]_i + [g_2]_i) = y_i]$. Without loss of generality, let's assume $[g]_i > 0$ (the proof is completely symmetric in the other case). Then, since $[g]_i$ is positive and g_2 has mean 0, we have

$\Pr[[g]_i + (g_2)_i < 0] \leq \frac{1}{2}$.

Moreover,

$$\begin{aligned}\Pr\left[[g]_i + [g_2]_i > 0\right] &= \Pr[[g_2]_i > 0]\Pr\left[[g]_i + [g_2]_i > 0 \mid [g_2]_i > 0\right] \\ &\quad + \Pr[[g_2]_i < 0]\Pr\left[[g]_i + [g_2]_i > 0 \mid [g_2]_i < 0\right]\end{aligned}$$

The first term is upper bounded by $\frac{1}{2}$ since $\Pr[[g_2]_i > 0] \leq \frac{1}{2}$. The second term we will bound using standard Gaussian tail bounds:

$$\begin{aligned}\Pr\left[[g]_i + [g_2]_i > 0 \mid [g_2]_i < 0\right] &\leq \Pr\left[|[g_2]_i| \leq |[g]_i| \mid [g_2]_i < 0\right] \\ &= \Pr[|[g_2]_i| \leq |[g]_i|] \leq \Pr[([g_2]_i)^2 \leq cD] = 1 - \Pr[([g_2]_i)^2 > cD] \\ &\leq 1 - \frac{2}{\sqrt{2\pi}}\exp\left(-cD/2\beta\right)\left(\sqrt{\frac{\beta}{cD}} - \left(\sqrt{\frac{\beta}{cD}}\right)^3\right)\end{aligned}$$

which implies

$$\Pr[[g_2]_i < 0]\Pr[[g]_i + [g_2]_i > 0 \mid [g_2]_i < 0] \leq \frac{1}{2}\left(1 - \frac{2}{\sqrt{2\pi}}\exp\left(-cD/2\beta\right)\left(\sqrt{\frac{\beta}{cD}} - \left(\sqrt{\frac{\beta}{cD}}\right)^3\right)\right)$$

Putting together, we have

$$\Pr[\text{sign}((g_1)_i + (g_2)_i) = y_i] \leq 1 - \frac{1}{\sqrt{2\pi}}\exp\left(-cD/2\beta\right)\left(\sqrt{\frac{\beta}{cD}} - \left(\sqrt{\frac{\beta}{cD}}\right)^3\right)$$

Together with the fact that $|\mathbb{S}| \geq \gamma n$ we get

$$\Pr[\mathbb{Y}_S = y_S | \mathbb{S} = s, g_1 = g] \leq \left[1 - \frac{1}{\sqrt{2\pi}}\exp\left(-cD/2\beta\right)\left(\sqrt{\frac{\beta}{cD}} - \left(\sqrt{\frac{\beta}{cD}}\right)^3\right)\right]^{\gamma n}$$

which implies that

$$H(\mathbb{Y}) \geq -\left(1 - \frac{1}{D}\right)\frac{(c-1)n}{c}\log\left[1 - \frac{1}{\sqrt{2\pi}}\exp\left(-cD/2\beta\right)\left(\sqrt{\frac{\beta}{cD}} - \left(\sqrt{\frac{\beta}{cD}}\right)^3\right)\right]$$

By setting $c = D = 3^{1/4}\sqrt{\beta}$ and a straightforward (albeit unpleasant) calculation, we can check that $H(\mathbb{Y}) \geq \frac{n}{25}\frac{(3^{1/4}\sqrt{\beta}-1)^2}{\sqrt{3\beta}}$, as we need.

□

We next show that the moments of the distribution are preserved up to a constant $\frac{\mathcal{G}}{1+\beta}$. The analysis is very

reminiscent of the standard analysis of Goemans-Williamson:

Lemma 133. *The distribution $\tilde{\mu}$ has $\frac{\mathcal{G}}{1+\beta}\Sigma_{i,j} \leq E_{\tilde{\mu}}[X_iX_j] \leq \frac{1}{1+\beta}\Sigma_{i,j}$*

Proof. Consider the Gram decomposition of $\tilde{\Sigma}_{i,j} = \langle v_i, v_j\rangle$. Then, $\mathcal{N}(0,\tilde{\Sigma})$ is in distribution equal to $(\text{sign}(\langle v_1, s\rangle), \ldots, \text{sign}(\langle v_n, s\rangle))$ where $s \sim \mathcal{N}(0, I)$. Similarly as in the analysis of Goemans-Williamson (Goemans and Williamson, 1995), if $\bar{v}_i = \frac{1}{\|v_i\|}v_i$, we have $\mathcal{G}\langle \bar{v}_i, \bar{v}_j\rangle \leq E_{\tilde{\mu}}[X_iX_j] = \frac{2}{\pi}\arcsin(\langle \bar{v}_i, \bar{v}_j\rangle) \leq \langle \bar{v}_i, \bar{v}_j\rangle$. However, since $\langle \bar{v}_i, \bar{v}_j\rangle = \frac{1}{\|v_i\|\|v_j\|}\langle v_i, v_j\rangle = \frac{1}{\|v_i\|\|v_j\|}\tilde{\Sigma}_{i,j} = \frac{1}{\|v_i\|\|v_j\|}\Sigma_{i,j}$ and $\|v_i\| = \sqrt{\tilde{\Sigma}_{i,i}} = \sqrt{1+\beta}, \forall i \in [1,n]$, we get that $\frac{\mathcal{G}}{1+\beta}\Sigma_{i,j} \leq E_{\tilde{\mu}}[X_iX_j] \leq \frac{1}{1+\beta}\Sigma_{i,j}$ as we want.

□

Lemma 132 and 133 together imply Theorem 130. Note that the above proof does not work in the general Ising model case: when $\tilde{\mathbb{E}}_{\tilde{q}}[x_ix_j]$ can be either positive or negative, even if we preserved each $\tilde{\mathbb{E}}_{\tilde{q}}[x_ix_j]$ up to a constant factor, this may not preserve the sum $\sum_{i,j} J_{i,j}\tilde{\mathbb{E}}_{\tilde{q}}[x_ix_j]$ due to cancellations in that expression.

In the next section, we will show how to handle this difficulty, and prove analogous results for general Ising models.

4.3.2 General Ising models

As noted in the previous section, the straightforward application of the rounding results in the previous section doesn't work, so we have to consider a different rounding – again inspired by roundings used in optimization.

The intuition is the same as in the ferromagnetic case: we wish to design a rounding which preserves the quadratic portion of the objective, while having a high entropy. In the previous section, this was achieved by modifying the Goemans-Williamson rounding so that it produces a high-entropy distribution. We will do a similar thing here, by modifying roundings due to (Charikar and Wirth, 2004) and (Alon et al., 2006).

The convex relaxation we will consider will be even simpler than the one before:

$$\max_{\tilde{q}\in \text{LA}(2)} \left\{ \sum_{i\sim j} J_{i,j}\tilde{\mathbb{E}}_{\tilde{q}}[x_ix_j] + n \right\} \tag{4.3.2}$$

We will prove the following two theorems:

Theorem 134 ((Li and Risteski, 2016))**.** *The relaxation (4.3.2) provides a factor $O(\log n)$ approximation to $\log Z$.*

Theorem 135 ((Li and Risteski, 2016))**.** *The relaxation (4.3.2) provides a factor $O(\log(\chi(G)))$ approximation to $\log Z$ where $\chi(G)$ is the chromatic number of G.*

As we mentioned before, since the chromatic number of a graph is bounded by n, the second theorem is in fact strictly stronger than the first, however the proof of the first theorem uses less heavy machinery, and is illuminating

enough to be presented on its own. Before delving into the proof of Theorem 134, we review the rounding used by (Charikar and Wirth, 2004) in the case of maximizing quadratic forms:

Algorithm 21 Quadratic form rounding by (Charikar and Wirth, 2004)

1: Input: A pseudo-moment matrix $\Sigma_{i,j} = \tilde{\mathbb{E}}_{\tilde{q}}[x_i x_j]$
2: Output: A sample x from a distribution ρ
3: Sample g from the standard Gaussian $N(0, I)$.
4: Consider the vector h, such that $h_i = g_i/T, T = \sqrt{4 \log n}$
5: Consider the vector r, such that $r_i = \frac{h_i}{|h_i|}$, if $|h_i| > 1$, and $r_i = h_i$ otherwise.
6: Produce the rounded vector $x \in \{-1, 1\}^n$, s.t.

$$x_i = \left\{ \begin{array}{ll} +1, & \text{with probability } \frac{1+r_i}{2} \\ -1, & \text{with probability } \frac{1-r_i}{2} \end{array} \right\}$$

Algorithm 22 Scaled down quadratic form rounding

1: Input: A pseudo-moment matrix $\Sigma_{i,j} = \mathbb{E}_{\nu}[x_i x_j]$
2: Output: A sample x from a distribution $\tilde{\mu}$
3: Sample g from the standard Gaussian $N(0, I)$.
4: Consider the vector h, such that $h_i = g_i/T, T = \sqrt{4 \log n}$
5: Consider the vector r, such that $r'_i = \frac{1}{2}\frac{h_i}{|h_i|}$, if $|h_i| > 1$, and $r'_i = \frac{1}{2}h_i$ otherwise.
6: Produce the rounded vector $x \in \{-1, 1\}^n$, s.t.

$$x_i = \left\{ \begin{array}{ll} +1, & \text{with probability } \frac{1+r_i}{2} \\ -1, & \text{with probability } \frac{1-r_i}{2} \end{array} \right\}$$

With that in hand, we can prove Theorem 134

Proof of Theorem 134. The proof again consists of exhibiting a rounding. Our rounding will essentially be the same as (Charikar and Wirth, 2004), except in step 3, we will produce a vector r'_i by scaling down the vector r_i by 2 coordinate-wise. For full clarity, the rounding is presented in Algorithm 22.

We again, need to analyze the entropy and the moments of the distribution $\tilde{\mu}$ that this rounding produces. Let us focus on the entropy first.

Since conditioning does not decrease entropy, it's true that $H(\tilde{\mu}) = H(x) \geq H(x|r)$, so it suffices to lower bound that quantity. However, note that it holds that $r_i \leq \frac{1}{2}$, and each x_i is rounded independently conditional on r_i, so we have:

$$H(x|r) = \sum_i H(x_i|r_i) = \sum_i \left(\frac{1+r_i}{2} \log\left(\frac{1+r_i}{2}\right) + \frac{1-r_i}{2}\left(\frac{1-r_i}{2}\right)\right) \geq \left(2 - \frac{3}{4}\log 3\right) n$$

Consider now the moments of the distribution.

Let us denote the distribution that the rounding 21 produces by ρ. By Theorem 1 in (Charikar and Wirth, 2004),

we have

$$\sum_{i,j} J_{i,j}\mathbb{E}_{\rho}[x_i x_j] \geq O\left(\frac{1}{\log n}\right)\sum_{i,j} J_{i,j}\mathbb{E}_{\nu}[x_i x_j]$$

Additional, both our and the (Charikar and Wirth, 2004) roundings are such that $\mathbb{E}_{\rho}[x_i x_j] = \mathbb{E}_{r}\mathbb{E}_{x|r}[x_i x_j]$ and $\mathbb{E}_{\tilde{\mu}}[x_i x_j] = \mathbb{E}_{r'}\mathbb{E}_{x|r'}[x_i x_j]$. Furthermore, as noted in (Charikar and Wirth, 2004), it is easy to check that $\mathbb{E}[x_i x_j|r'] = r'_i r'_j$ and obviously $r'_i = 2r_i, \forall i$ in distribution, so we have:

$$\mathbb{E}_{\tilde{\mu}}[x_i x_j] = \mathbb{E}_{r'}\mathbb{E}_{\mathbf{x}|r'}[x_i x_j] = \frac{1}{4}\mathbb{E}_{r}\mathbb{E}_{x|r}[x_i x_j] = \frac{1}{4}\mathbb{E}_{\rho}[x_i x_j]$$

But, this directly implies

$$\sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}[x_i x_j] = \frac{1}{4}\sum_{i,j} J_{i,j}\mathbb{E}_{\rho}[x_i x_j] \geq O\left(\frac{1}{\log n}\right)\sum_{i,j} J_{i,j}\mathbb{E}_{\nu}[x_i x_j]$$

as we needed. □

Next, we prove the more general Theorem 135.

Before proceeding, let's recall for completeness the following definition of a chromatic number.

Definition (Chromatic number)**.** The chromatic number $\chi(G)$ of a graph $G = (V(G), E(G))$ is defined as the minimum number of colors in a coloring of the vertices $V(G)$, such that no vertices $i, j : (i, j) \in E(G)$ are colored with the same color.

Also, let us denote by $\mathcal{S}^{n-1}$ the set of unit vectors in $\mathbb{R}^n$ and $L_\infty[0, 1]$ the set of (essentially) bounded functions: the functions which are bounded except on a set of measure zero.

Then, we can recall Theorem 3.3 from (Alon et al., 2006):

Theorem 136 ((Alon et al., 2006))**.** *There exists an absolute constant c such that the following holds: Let $G = (V(G), E(G))$ be an undirected graph on n vertices without self-loops*[1]*, let $\chi(G)$ be the chromatic number of G. Then for every function $f : V(G) \to \mathcal{S}^{n-1}$, there exists a function $F : V \to L_\infty[0, 1]$ so that for every $i \in V(G)$, $\|F(i)\|_\infty \leq \sqrt{c\chi(G)}$ and for every $(i, j) \in E(G)$,*

$$\langle f(i), f(j)\rangle = \int_0^1 F(i)(t)F(j)(t)dt$$

Now, we can prove Theorem 135

Proof of Theorem 135. The proof is similar, though a little more complicated than the proof of Theorem 134.

Let $\tilde{\mathbb{E}}_{\nu}[\cdot]$ be the solution of the relaxation. By matrix formulation of the pseudo-moment relaxation in Section **??** , we know that $\tilde{\mathbb{E}}_{\nu}[x_i x_j] = \langle f(i), f(j)\rangle$ for some unit vectors $f(i), f(j)$.

[1] Meaning no edge connects a vertex with itself

Hence, by theorem 136, there exists a function $F : V \to L_\infty[0,1]$ so that for every $i \in V(G)$, $\|F(i)\|_\infty \leq \sqrt{c\chi(G)}$ and for every $(i,j) \in E(G)$,

$$\tilde{\mathbb{E}}_v[x_i x_j] = \int_0^1 F(i)(t)F(j)(t)dt$$

Consider the following rounding:

- Pick a t uniformly at random from $[0,1]$.
- Consider the function $h_t : V \to R$, such that $h_t(i) = \frac{F(i)(t)}{2\sqrt{c\chi(G)}}$
- Produce the rounded vector $\mathbf{x} \in \{-1,1\}^{V(G)}$, s.t.

$$x_i = \left\{ \begin{array}{ll} +1, & \text{with probability } \frac{1+h_t(i)}{2} \\ -1, & \text{with probability } \frac{1-h_t(i)}{2} \end{array} \right\}$$

Note importantly that the algorithm does not need to perform this rounding – it is for the analysis of the approximation factor of the relaxation. Therefore, we need not construct it algorithmically.

Let us denote this distribution as $\tilde{\mu}$. We first show that $\tilde{\mu}$ has entropy at least $\left(2 - \frac{3}{4}\log 3\right)n$. Note that each x_i are round independently conditional on t. Moreover, since $\|F(v)\|_\infty \leq \sqrt{c\chi(G)}$, we know that $h_t(v) \leq \frac{1}{2}$. Therefore, for every fixed $t_0 \in [0,1]$

$$\begin{aligned} H(\tilde{\mu} \mid t = t_0) &= \sum_{i \in V(G)} H(x_i \mid t = t_0) \\ &= \sum_{i \in V(G)} \left(\frac{1+h_{t_0}(v)}{2} \log \frac{1+h_{t_0}(v)}{2} + \frac{1-h_{t_0}(v)}{2} \log \frac{1-h_{t_0}(v)}{2} \right) \\ &\geq \left(2 - \frac{3}{4}\log 3\right)n \end{aligned}$$

Integrating over t_0 we get that $H(\tilde{\mu}) \geq \left(2 - \frac{3}{4}\log 3\right)n$.

Next, we will show that $\tilde{\mu}$ preserves the "energy" part of the objective up to a multiplicative factor $O(\log \chi(G))$: Consider each edge $(i,j) \in E(G)$. We have:

$$\mathbb{E}_{\tilde{\mu}}[x_i x_j] =$$

$$\int_0^1 \left(\frac{(1+h_t(i))(1+h_t(j))}{4} + \frac{(1-h_t(i))(1-h_t(j))}{4} - \frac{(1+h_t(i))(1-h_t(j))}{4} - \frac{(1-h_t(i))(1+h_t(j))}{4} \right) dt$$

$$= \int_0^1 h_t(i)h_t(j)dt = \frac{1}{4c\chi(G)} \int_0^1 F(i)(t)F(j)(t)dt = \frac{1}{4c\chi(G)} \tilde{\mathbb{E}}_v[x_i x_j]$$

This implies that

$$\sum_{i,j\in E(G)} J_{i,j}\mathbb{E}_{\tilde{\mu}}[x_i x_j] \geq \frac{1}{4c\chi(G)} \sum_{i,j\in E(G)} J_{i,j}\tilde{\mathbb{E}}_{\nu}[x_i x_j]$$

Therefore, the relaxation provides a factor $O(\chi(G))$ approximation of $\log \mathcal{Z}$, as we wanted.

□

4.4 Guarantees for dense and low threshold-rank Ising models using entropy-respecting roundings

In this section, we will show how we can leverage well-design entropy approximations $\tilde{H}$ in structured instances. More precisely, we will be considering *dense* and *low threshold-rank* Ising models. These terms might be familiar to those who have studied combinatorial optimization in the context of constraint satisfaction problems. We will generalize these concepts to Ising models in the obvious manner.

An Ising model $p(x) \propto \exp\left(\sum_{i\sim j} J_{i,j}x_i x_j\right), x \in \{-1,1\}^n$ is Δ-*dense* if it satisfies $\Delta|J_{i,j}| \leq \frac{J_T}{n^2}, \forall i,j \in [n]$, where $J_T = \sum_{i,j}|J_{i,j}|$.

This is a natural generalization of the typical way to define density for combinatorial optimization problems (see e.g. (Yoshida and Zhou, 2014)). To see this consider a graph $G = (V,E)$ with $|E| = cn^2$. For constraint satisfaction problems, we care about objectives that look like

$$\mathbb{E}_{e\in E} f(e) = \sum_{e\in E} \frac{1}{|E|} f(e)$$

for some function f. Hence, the "weight" in front of each pair (i,j) in the objective is 0 if there is no edge or $\frac{1}{|E|}$. This corresponds to $\Delta = \frac{1}{c}$ in our definition. For partition function problems, however, scale matters (i.e. we cannot assume $\sum_{i,j} J_{i,j} = 1$), so the above generalization is very organic.

We will then show:

Theorem 137 ((Risteski, 2016))**.** *For Δ-dense Ising models, there is an algorithm based on Sherali-Adams hierarchies which achieves an additive approximation of ϵJ_T to $\log Z$, where $Z = \sum_{x\in\{-1,1\}^n} \exp\left(\sum_{i\sim j} J_{i,j}x_i x_j\right)$ and runs in time $n^{O\left(\frac{1}{\Delta\epsilon^2}\right)}$.*

Our second contribution are analogous claims for Ising models whose potentials look like low rank matrices. (More precisely, adjacency matrices of *low threshold rank* graphs, a concept introduced by (Arora et al., 2010) in the context of their algorithm for Unique Games.)

Concretely, an Ising model $p(x) \propto \exp\left(\sum_{i\sim j} J_{i,j}x_i x_j\right), x \in \{-1,1\}^n$ is *regular* if $\sum_{j:i\sim j}|J_{i,j}| = J', \forall i$. The *adjacency matrix* of a regular Ising model is the matrix $A_{i,j} = |J_{i,j}|/J'$. Then, we show:

Theorem 138 ((Risteski, 2016)). *There is an algorithm based on Lasserre hierarchies which achieves an additive aproximation of $\epsilon n J'$ to $\log Z$, where $Z = \sum_{x\in\{-1,1\}^n} \exp\left(\sum_{i\sim j} J_{i,j} x_i x_j\right)$, and runs in time $n^{rank(\Omega(\epsilon^2))/\Omega(\epsilon^2)}$, where $rank(\tau)$ is the number of eigenvalues of the adjacency matrix A greater than or equal to τ.*

Interestingly this property of the graph was previously introduced for purposes of combinatorial optimization problems like small-set expansion, Unique Games (Steurer, 2010; Arora et al., 2010) – yet it also helps with counting type problems. Note that since we prove additive factor guarantees to $\log Z$, using the fact the $e^\epsilon \leq 1 + 2\epsilon$, we can easily turn them to multiplicative factor guarantees on Z.

While these guarantees are not as strong as one usually gets in the correlation decay regime (i.e. $1+\epsilon$ multiplicative factor approximations to Z in time poly$(n, \frac{1}{\epsilon})$), to the best of our knowledge, these are the first approximations guarantees for Z when correlation decay does not hold. We discuss interesting regimes of the potentials $J_{i,j}$ in Section 4.4.4.

4.4.1 Entropy-respecting roundings

In contrast to Section4.3, we will use more carefully designed entropy approximations. The rounding of the corresponding relaxation will be such that it *exactly* preserves the entropy approximation we design – so the task will be to analyze how well we preserve the quadratic portion of the objective. Hence, we call them *entropy-preserving* roundings.

Recall, $\tilde{H}$ needs to be an upper bound on the entropy of a distribution on which we have essentially no handle other than having the first few moments through the convex programming hierarchies we are using. A natural candidate to get an upper bound is the *chain rule*.

Towards that, notice that for any set S of size at most k, where k is the number of levels of the Sherali-Adams or Lasserre hierarchy, $H(\mu_S)$ is a well-defined quantity: it's exactly

$$H(\mu_S) = \sum_{x_S \in \{-1,1\}^{|S|}} \mu_S(x_S) \log(\mu_s(x_S))$$

Since these local quantities are essentially all the information about the joint distribution μ we have, our functional must involve such quantities only.

The simplest functional one can design surely is the following:

Definition. The *mean-field pseudo-entropy functional* $H_{\text{MF}}(\mu)$ is defined as $H_{\text{MF}}(\mu) = \sum_{i=1}^n H(\mu_i)$.

Remark. Note, this is *not* the same as the usual mean-field approximation in statistical physics. The mathematical program analogue of that approximation would be to enforce that $\mathbb{E}_\mu[x_i x_j] = \mathbb{E}_\mu[x_i]\mathbb{E}_\mu[x_j]$ – which would result in a non-convex relaxation generally. We think the name is appropriate though, since the bound on the entropy is *mean-field*, i.e. results by treating μ as if it were a product distribution.

Almost trivially for any $\mu \in \mathcal{M}$, the following proposition holds:

Proposition 139. *For any distribution* $\mu : \{-1,1\}^n \to [0,1]$, $H(\mu) \leq H_{MF}(\mu)$

Proof. By the chain rule, $H(\mu) = \sum_{i=1}^{n} H(\mu_i|\mu_{[i-1]})$, where $[i-1]$ denotes the set $\{1,2,\ldots,i-1\}$ and $H(X|Y)$ is the conditional entropy of X given Y. However, since $H(\mu_i|\mu_{[i-1]}) \leq H(\mu_i)$ the claim trivially holds. □

We will also consider generalizations of the above – where before applying the above "mean-field" bound on the entropy, one can condition on a small subset first. Namely,

Definition. The *augmented mean-field pseudo-entropy functional* for subsets of size k, $H_{\mathrm{aMF},k}(\mu)$ is defined as $H_{\mathrm{aMF},k}(\mu) = \min_{|S|\leq k}\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\}$.

The same proof as in Proposition 139 implies:

Proposition 140. $H(\mu) \leq H_{aMF,k}(\mu)$

Furthermore, it's quite easy to show that $H_{\mathrm{aMF},k}(\mu)$, like $H_{\mathrm{MF}}(\mu)$, is a concave function.

Lemma 141. *The pseudo-entropy functional* $H_{aMF,k}(\mu) = \min_{|S|\leq k}\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\}$ *is concave in the variables* $\{\mu_{S\cup\{i\}}(x_{S\cup\{i\}}) \,||S| \leq k, i \in [n]\}$.

Proof. Since $H_{\mathrm{aMF},k}(\mu) = \min_{|S|\leq k}\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\}$, and the minimum of concave functions is concave, all we need to show is that $H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)$ is concave for all S. It's well known that entropy is a concave function, so $H(\mu_S)$ is concave. What remains to be shown is that $\sum_{i\notin S} H(\mu_i|\mu_S)$ is concave. But, since the sum of concave functions is concave, it suffices to prove $H(\mu_i|\mu_S)$ is concave.

The proof of this is essentially the same as the proof of concavity of entropy. Abusing notation a bit, we will denote as $\mu_A|x_B$ the conditional distribution on the variables in A, conditioned on the variables in B having the value x_B. We recall that

$$\begin{aligned}
H(\mu_i|\mu_S) &= \sum_{x_S\in\{-1,1\}^{|S|}} \mu_s(x_S) H(\mu_i|x_s) \\
&= -\sum_{x_S\in\{-1,1\}^{|S|}} \sum_{x_i\in\{-1,1\}} \mu_s(x_S)\mu_{i|x_S}(x_i)\log(\mu_{i|x_S}(x_i)) \\
&= -\sum_{x_S\in\{-1,1\}^{|S|}} \sum_{x_i\in\{-1,1\}} \mu_{S\cup\{i\}}(x_{S\cup\{i\}})\log(\mu_{i|x_S}(x_i)) \\
&= -\sum_{x_S\in\{-1,1\}^{|S|}} \sum_{x_i\in\{-1,1\}} \mu_{S\cup\{i\}}(x_{S\cup\{i\}})\log\left(\frac{\mu_{S\cup\{i\}}(x_{S\cup\{i\}})}{\mu_s(x_S)}\right)
\end{aligned}$$

We rewrite the last expression as a KL divergence as follows:

$$-\sum_{x_S\in\{-1,1\}^{|S|}} \sum_{x_i\in\{-1,1\}} \mu_{S\cup\{i\}}(x_{S\cup\{i\}})\log\left(\frac{\mu_{S\cup\{i\}}(x_{S\cup\{i\}})}{\mu_s(x_S)\frac{1}{2}}\right) + 1 = -KL(\mu_{S\cup\{i\}}||(\mu_S \times r)) + 1 \tag{4.4.1}$$

where r is a uniform distribution over $\{-1, 1\}$.

Then, if $\mu^{\lambda}_{S\cup\{i\}} = \lambda\mu^{1}_{S\cup\{i\}} + (1-\lambda)\mu^{2}_{S\cup\{i\}}$, we want to show

$$H(\mu_i^{\lambda}|\mu_S^{\lambda}) \geq \lambda H(\mu_i^1|\mu_S^1) + (1-\lambda)H(\mu_i^2|\mu_S^2)$$

By (4.4.1) and the convexity of KL divergence,

$$\begin{aligned} H(\mu_i^{\lambda}|\mu_S^{\lambda}) &= -KL(\mu^{\lambda}_{S\cup\{i\}}\|(\mu_S^{\lambda}\times r)) + 1 \\ &\geq -\lambda KL(\mu^{1}_{S\cup\{i\}}\|(\mu_S^{1}\times r)) - (1-\lambda)KL(\mu^{2}_{S\cup\{i\}}\|(\mu_S^{2}\times r)) + 1 \\ &= \lambda H(\mu_i^1|\mu_S^1) + (1-\lambda)H(\mu_i^2|\mu_S^2) \end{aligned}$$

which is what we want.

□

4.4.2 Guarantees for dense Ising models using entropy-respecting roundings

We will now use the entropy approximations roundings to design provable algorithms for approximating partition functions of dense Ising models.

There are multiple reasons to study this particular subclass: from the theoretical computer science point of view, we have various PTAS for constraint satisfaction problem when the constraint graph is dense (Yoshida and Zhou, 2014; Arora et al., 1995) so we might hope to get results better than the worst-case one ones for partition function calculation as well.

Another motivation comes from *mean-field* ferromagnetic Ising model (also known as the *Curie-Weiss* model (Ellis and Newman, 1978)), which is frequently studied as a very simplified model of ferromagnetism because one can get relatively easily results about global properties of the model like the partition function, magnetization, etc. In the mean-field model, each spin interacts (equally strongly) with every other spin.

We will, in this section, generalize the classical results about the ferromagnetic Curie-Weiss model, as well as provide the natural counterpart of the results in (Yoshida and Zhou, 2014; Arora et al., 1995) for partition functions.

Let us first review the standard results about Curie-Weiss. Recall, this model follows the distribution $p(x) \propto \exp\left(\sum_{i,j=1}^{n} \frac{J}{n} x_i x_j\right)$, $J > 0$. It is easy to analyze because $p(x)$ factorizes and can be "reparametrized" in terms of the magnetization. Namely, since $\sum_{i,j=1}^{n} \frac{J}{n} x_i x_j = \frac{J}{n}(\sum_i x_i)^2$, and $(\sum_i x_i)^2 \in [-n, n]$, one can show (Ellis and Newman, 1978):

Theorem 142 ((Ellis and Newman, 1978)). *For the Curie-Weiss model,*

$$\log \mathcal{Z} = (1 \pm o(1))\left(n \max_{m\in[-1,1]}\left(Jm^2 + \frac{1-m}{2}\log\frac{1-m}{2} + \frac{1+m}{2}\log\frac{1+m}{2}\right)\right)$$

The proof of this theorem involves rewriting the expression for Z as follows:

$$Z = \sum_{x\in\{-1,1\}^n} \exp\left(\sum_{i,j}\frac{J}{n}x_i x_j\right) = \sum_l \exp\left(\frac{J}{n}l^2\right)\cdot n_l$$

where n_l is the number of terms where $\sum_{i=1}^n x_i = l$. Then, using Stirling's formula and some more algebraic manipulation, one can estimate the dominating term in the summation. The claim of the theorem then follows.

We significantly generalize the above claim using notions from theoretical computer science. The goal is to prove Theorem 137.

Let $J_T = \sum_{i,j}|J_{i,j}|$. As discussed in Section 4.4, we define the following notion of density inspired by the definition of a dense graph in combinatorial optimization (Yoshida and Zhou, 2014):

Definition. An Ising model is Δ-*dense* if $\forall i \neq j, \Delta|J_{i,j}| \leq \frac{J_T}{n^2}$, $\Delta \in (0,1]$.

We will consider the relaxation for $\log Z$ given by the augmented pseudo-entropy functional and the level $k = O(1/(\Delta\epsilon^2))$ Sherali-Adams relaxation, namely:

$$\max_{\mu\in \mathrm{SA}(k), k=O(1/(\Delta\epsilon^2))}\left\{\sum_{i,j}J_{i,j}\mathbb{E}_\mu\left[x_i x_j\right] + H_{\mathrm{aMF},k}(\mu)\right\} \tag{4.4.2}$$

We also recall *correlation rounding* as defined in (Barak et al., 2011). In correlation rounding, we pick a "seed set" of a certain size, condition on it, and round the rest of the variables independently. The usual thing to prove is that there is a good "seet set" of a small size to condition on. In particular, for the dense case, the following lemma was proven in (Yoshida and Zhou, 2014):

Lemma 143 ((Yoshida and Zhou, 2014)). *There exists a set S of size $k = O(1/(\Delta\epsilon^2))$, s.t.*

$$\left|\sum_{i,j}J_{i,j}\mathbb{E}_\mu\left[x_i x_j|x_S\right] - \sum_{i,j}J_{i,j}\mathbb{E}_\mu\left[x_i|x_S\right]\mathbb{E}_\mu\left[x_j|x_S\right]\right| \leq \frac{100}{\Delta k}J_T$$

With this in hand, we proceed to the main theorem of this section:

Theorem 144 (Restatement of Theorem 137). *The output of 4.4.2 is an ϵJ_T additive approximation to* $\log Z$.

Proof. The function 4.4.2 is optimizing is a sum of two terms: $\sum_{i\sim j} J_{i,j}\mathbb{E}_\mu\left[x_i x_j\right]$ and an entropy term. Following standard terminology in statistical physics, we will call the former term *average energy*.

We will analyze the quality of the convex relaxation by exhibiting a *rounding* of the pseudo-distribution to an actual distribution. There is a difference in what this means compared to the roundings we use in combinatorial optimization: there we only care about producing a *single* $\{+1,-1\}$ solution. Here, because of the entropy term, it's essential that we produce a *distribution* over $\{+1,-1\}$ solutions.

We use the fact that correlation rounding can be viewed as producing distributions with a fairly explicit expression for their entropy. Let S be the set of size $O(\frac{1}{\Delta\epsilon^2})$ that Lemma 143 gives. Consider the distribution $\tilde{\mu}(x) = \mu(x_S)\Pi_{i\notin S}\mu(x_i|x_S)$[2]. In other words, this is the distribution which rounds the variables in S according to their local distribution, and all other variables independently according to the conditional distribution on $\mathbf{x}_S$.

Consider the average energy first. By Lemma 143,

$$\left|\sum_{i,j} J_{i,j}\mathbb{E}_\mu\left[x_i x_j|x_S\right] - \sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}\left[x_i x_j|x_S\right]\right| \leq J_T\epsilon$$

Now consider the entropy term. The entropy of the distribution $\tilde{\mu}$ is $H(\tilde{\mu}) = H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)$. But, since $H_{\text{aMF},k}(\mu) = \min_{|S|\leq k}\{H(\mu_S) + \sum_{i\notin S} H(\mu_i|\mu_S)\}$, $H_{\text{aMF},k}(\mu) \leq H(\tilde{\mu})$ follows. This immediately implies that

$$\left(\sum_{i,j} J_{i,j}\mathbb{E}_\mu\left[x_i x_j\right] + H_{\text{aMF},k}(\mu)\right) - \left(\sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}\left[x_i x_j\right] + H(\tilde{\mu})\right) =$$

$$\left(\sum_{i,j} J_{i,j}\mathbb{E}_\mu\left[x_i x_j\right] - \sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}\left[x_i x_j\right]\right) + (H_{\text{aMF},k}(\mu) - H(\tilde{\mu})) \leq J_T\epsilon$$

This exactly proves the claim we want.

□

Notice, in the case of the Curie-Weiss model, since $J > 0$, the value of the relaxation 4.4.2 is at least J_T, Theorem 144 gives a $1+\epsilon$ multiplicative factor approximation to $\log Z$ for any constant ϵ, so generalizes the statement of Theorem 142 to cases where the potentials $J_{i,j}$ might vary in magnitude and sign.

4.4.3 Guarantees for low threshold rank Ising models using entropy-respecting Ising models

Concretely, an Ising model $p(x) \propto \exp\left(\sum_{i\sim j} J_{i,j}x_i x_j\right), x \in \{-1,1\}^n$ is *regular* if $\sum_{j:i\sim j}|J_{i,j}| = J', \forall i$. The *adjacency matrix* of a regular Ising model is the matrix $A_{i,j} = |J_{i,j}|/J'$. Then, we show:

[2]Notice this is an actual, well-defined distribution, and not only a pseudo-distribution anymore.

Theorem 145. *There is an algorithm based on Lasserre hierarchies which achieves an additive aproximation of $\epsilon n J'$ to $\log Z$, where $Z = \sum_{x\in\{-1,1\}^n} \exp\left(\sum_{i\sim j} J_{i,j}x_i x_j\right)$, and runs in time $n^{rank(\Omega(\epsilon^2))/\Omega(\epsilon^2)}$, where $rank(\tau)$ is the number of eigenvalues of the adjacency matrix A greater than or equal to τ.*

Interestingly this property of the graph was previously introduced for purposes of combinatorial optimization problems like small-set expansion, Unique Games (Steurer, 2010; Arora et al., 2010) – yet it also helps with counting type problems. Note that since we prove additive factor guarantees to $\log Z$, using the fact the $e^{\epsilon} \leq 1 + 2\epsilon$, we can easily turn them to multiplicative factor guarantees on Z.

If we use the added power of the Lasserre hierarchy, we can also handle Ising models whose weights look like low rank matrices. We want to prove Theorem 145.

We will consider for simplicity in this section *regular* Ising models in the weighted sense, meaning $\sum_j |J_{i,j}| = J', \forall i$ [3]. The *adjacency matrix* of an Ising model will be the doubly-stochastic matrix with entries $|J_{i,j}|/J'$.

Let's recall the definition of threshold rank from (Arora et al., 2010):

Definition. The τ-threshold rank of a regular graph is the number of eigenvalues of the normalized adjacency matrix greater than or equal to τ.

We will, in analogy, define the threshold rank of an Ising model.

Definition. The τ-threshold rank of a regular Ising model is the number of eigenvalues of its adjacency matrix greater than or equal to τ.

We will consider the following convex program:

$$\max_{\mu\in \mathrm{LAS}(k)} \left\{ \sum_{i,j} J_{i,j}\mathbb{E}_{\mu}\left[x_i x_j\right] + H_{\mathrm{aMF},k}(\mu) \right\} \tag{4.4.3}$$

Consider the vectors $v_i, i \in [n]$, s.t. $\langle v_j, v_j\rangle = \mathbb{E}_{\mu}\left[x_i x_j\right]$. Then, (Barak et al., 2011) prove that when the graph has low threshold rank, "local" correlations propagate to "global" correlations, and as a consequence of this, there is a set of size at most $\mathrm{rank}(\Omega(\epsilon^2))/\Omega(\epsilon^2)$, such that conditioning on it causes the $\left|\sum_{i,j} J_{i,j}\mathbb{E}_{\mu}\left[x_i x_j|x_S\right] - \sum_{i,j} J_{i,j}\mathbb{E}_{\tilde{\mu}}\left[x_i x_j|x_S\right]\right|$ to drop below ϵJ_T. More precisely:

Lemma 146 ((Barak et al., 2011))**.** *There exists a set S of size $t \leq rank(\Omega(\epsilon^2))/\Omega(\epsilon^2)$, where $rank(\tau)$ is the τ-threshold rank of the Ising model, s.t.* [4]

$$\left|\sum_{i,j} J_{i,j}\mathbb{E}_{\mu}\left[x_i x_j|x_S\right] - \sum_{i,j} J_{i,j}\mathbb{E}_{\mu}\left[x_i|x_S\right]\mathbb{E}_{\mu}\left[x_j|x_S\right]\right| \leq \epsilon J_T$$

[3] Though we remind again, all of the claims can be appropriately generalized at the expense of more bothersome notation.
[4] Note, $J_T = nJ'$ in this case.

Hence, analogously as in Theorem 144, we get:

Theorem 147 (Restatement of Theorem 145)**.** *The output of 4.4.3 is a ϵJ_T additive approximation to* $\log Z$.

4.4.4 Discussion on interpreting the results

Finally, we comment briefly on how to interpret the results in this section in different "temperature regimes" i.e. different scales of the potentials $J_{i,j}$. Note that partition function approximation problems are not scale-invariant, and their hardness is sensitive to the size of the coefficients $J_{i,j}$.

For simplicity of the discussion, let's assume the graph G is d-regular. There are generically three regimes for the problem:

- "High temperature regime", i.e. when $|J| = O\left(\frac{1}{d}\right)$ for a sufficiently small constant in the $O(\cdot)$ notation. In this case, standard techniques like Dobrushin's uniqueness criterion show that there is correlation decay. This is the regime where generically Markov Chain methods work. Note that using such methods, generally one can get a $(1+\epsilon)$-factor approximation for Z in time $\text{poly}\left(n, \frac{1}{\epsilon}\right)$, which is unfortunately much stronger than what our method gets in that regime. It would be extremely interesting to see if the methods in our paper can be modified to subsume this regime as well.

- "Around the transition threshold", i.e. when $|J| = \Theta(\frac{1}{d})$ for a sufficiently large constant in the Θ notation, such that there is no correlation decay. Generally, unless there is some special structure, Markov Chain methods will provide *no non-trivial* guarantee in this regime – however, we get an order ϵn additive factor approximation to $\log Z$, which translates to a $(1+\epsilon)^n$ factor approximation of Z. We do not, to the best of our knowledge, know how to get such results using *any other methods.*

- "Low temperature regime", i.e. when $|J| = \omega(1/d)$. In this case, in light of the variational characterization of $\log Z$ and the fact that the entropy is upper bounded by n, the dominating term will typically be the energy term $\sum_{(i,j)\in E(G)} J_{i,j}\mathbb{E}_\mu[x_i x_j]$, so essentially the quality of approximation will be dictated by the hardness of the optimization problem corresponding to the energy term. (e.g., for the anti-ferromagnetic case, where all the potentials $J_{i,j}$ are negative, the optimization problem corresponding to the energy term is just max-cut, and we cannot hope for more than a constant factor approximation to $\log Z$ for general (negative) potentials.)

Appendix A

Notations

We will use standard linear algebra and probability notation.

With respect to vector notation, we denote by $\mathbf{0}$ the all-zeroes vector and $\mathbf{1}$ the all-ones vector. We write $e_i \in \mathbb{R}^d, i \in [d]$ for the i-th canonical vector. We denote by $\mathrm{supp}(x)$ the support of a vector (or matrix) x.

With respect to matrix notation, A^+ will denote the Moore-Penrose pseudo-inverse of a matrix A, and for symmetric matrices A, we use $A^{-1/2}$ as a shorthand for $(A^+)^{1/2}$. The least *non-zero* singular value of matrix A is denoted $\sigma_{\min}(A)$, and the largest $\sigma_{\max}(A)$. A_i will denote the i-th column of matrix A, and A^i the i-th row. We use $\exp(A)$ for the matrix or vector obtained by taking the *entries-wise* exponential. For matrices A, B we define the Kronecker product $\otimes$ as $(A \otimes B)_{ijkl} = A_{ij}B_{kl}$. A useful identity is that $(A \otimes B) \cdot (C \otimes D) = (AC) \otimes (BD)$ whenever the matrix multiplications are defined.

With respect to asymptotic notation, we write $a \lesssim b$ if there exists a universal constant c such that $a \leq cb$ and we define $\gtrsim$ similarly. An analogous definition applies to the PSD ordering of matrices. The notation $\tilde{O}(\cdot)$ hides polylogarithmic factors.

Bibliography

A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, 2013.

Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006.

Noga Alon, Konstantin Makarychev, Yury Makarychev, and Assaf Naor. Quadratic forms on graphs. *Inventiones mathematicae*, 163(3):499–522, 2006.

A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade. Learning latent bayesian networks and topic models under expansion constraints. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

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(1):2773–2832, 2014.

S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond svd. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2012a.

S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013a.

S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, 2014.

S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, 2015a.

Sanjeev Arora, David Karger, and Marek Karpinski. Polynomial time approximation schemes for dense instances of np-hard problems. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pages 284–293. ACM, 1995.

Sanjeev Arora, Boaz Barak, and David Steurer. Subexponential algorithms for unique games and related problems. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 563–572. IEEE, 2010.

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, 2012b.

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, 2012c.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013b.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. 2015b.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski. New practical algorithms for learning noisy or networks via symmetric nmf. 2017a.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski. Provable learning of noisy-or networks. In *Symposium on the Theory of Computing (STOC)*, 2017b.

Pranjal Awasthi and Andrej Risteski. On some provably correct cases of variational inference for topic models. In *Advances in Neural Information Processing Systems*, pages 2098–2106, 2015.

T. Bansal, C. Bhattacharyya, and R. Kannan. A provable svd-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Boaz Barak, Prasad Raghavendra, and David Steurer. Rounding semidefinite programming hierarchies via global correlation. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 472–481. IEEE, 2011.

Boaz Barak, Jonathan A Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 31–40. ACM, 2014.

D. Blei, A. Ng, , and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

David M Blei and John D Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34, 2009.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

Venkat Chandrasekaran, Misha Chertkov, David Gamarnik, Devavrat Shah, and Jinwoo Shin. Counting independent sets using the bethe approximation. *SIAM Journal on Discrete Mathematics*, 25(2):1012–1034, 2011.

Moses Charikar and Anthony Wirth. Maximizing quadratic programs: extending grothendieck's inequality. In *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, pages 54–60. IEEE, 2004.

Michael Chertkov and Vladimir Y Chernyak. Loop series for discrete statistical models on graphs. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(06):P06009, 2006.

David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

Imre Csiszśar and Gábor Tusnády. Information geometry and alternating minimization procedures. In *Statistics and Decisions, Supplemental Issue Number 1*, 1984.

S. Dasgupta and L. Schulman. A two-round variant of em for gaussian mixtures. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2000.

S. Dasgupta and L. Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *Journal of Machine Learning Research*, 8:203–226, 2007.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

W. Ding, M.H. Rohban, P. Ishwar, and V. Saligrama. Topic discovery through data dependent and random projections. *arXiv preprint arXiv:1303.3664*, 2013.

W. Ding, M.H. Rohban, P. Ishwar, and V. Saligrama. Efficient distributed topic modeling with provable guarantees. In *Proceedings ot the 17th International Conference on Artificial Intelligence and Statistics*, pages 167–175, 2014.

Richard S Ellis. *Entropy, large deviations, and statistical mechanics*, volume 271. Springer Science & Business Media, 2012.

Richard S Ellis and Charles M Newman. The statistics of curie-weiss models. *Journal of Statistical Physics*, 19(2): 149–161, 1978.

Sacha Friedli and Yvan Velenik. *Statistical Mechanics of Lattice Systems: a Concrete Mathematical Introduction*. Cambridge University Press, 2017.

Rong Ge, Qingqing Huang, and Sham M Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 761–770. ACM, 2015.

Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

Robert B Griffiths. Correlations in ising ferromagnets. i. *Journal of Mathematical Physics*, 8(3):478–483, 1967.

Yonatan Halpern and David Sontag. Unsupervised learning of noisy-or bayesian networks. *arXiv preprint arXiv:1309.6834*, 2013.

Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

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.

Ilse CF Ipsen. Relative perturbation results for matrix eigenvalues and singular values. *Acta numerica*, 7:151–201, 1998.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Yacine Jernite, Yonatan Halpern, and David Sontag. Discovering hidden variables in noisy-or networks using quartet tests. In *Advances in Neural Information Processing Systems*, pages 2355–2363, 2013.

Mark R Jerrum, Leslie G Valiant, and Vijay V Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:169–188, 1986.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

A. Kumar and R. Kannan. Clustering with spectral norm and the k-means algorithm. In *Proceedings of Foundations of Computer Science (FOCS)*, 2010.

Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

D. Lee and S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2000.

David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

Ren-Cang Li. Relative perturbation theory. iii. more bounds on eigenvalue variation. *Linear algebra and its applications*, 266:337–345, 1997.

Ren-Cang Li. Relative perturbation theory: I. eigenvalue and singular value variations. *SIAM Journal on Matrix Analysis and Applications*, 19(4):956–982, 1998a.

Ren-Cang Li. Relative perturbation theory: Ii. eigenspace and singular subspace variations. *SIAM Journal on Matrix Analysis and Applications*, 20(2):471–492, 1998b.

Yuanzhi Li and Andrej Risteski. Approximate maximum entropy principles via goemans-williamson with applications to provable variational methods. In *30th Annual Conference on Neural Information Processing Systems (NIPS)*, 2016.

Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of non-negative matrix factorization via alternating updates. In *Advances in Neural Information Processing Systems*, 2016.

Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 438–446. IEEE, 2016.

Ofer Meshi, Ariel Jaimovich, Amir Globerson, and Nir Friedman. Convexifying the bethe free energy. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 402–410. AUAI Press, 2009.

Randolph A Miller, Harry E Pople Jr, and Jack D Myers. Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307(8):468–476, 1982.

Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375. ACM, 2005.

P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

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

Andrej Risteski. How to calculate partition functions using convex programming hierarchies: provable bounds for variational methods. In *29th Annual Conference on Learning Theory (COLT)*, pages 1402–1416, 2016.

Michael Shwe and Gregory Cooper. An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network. *Computers and Biomedical Research*, 24(5):453–475, 1991.

Tomas Singliar and Milos Hauskrecht. Noisy-or component analysis and its application to link analysis. *Journal of Machine Learning Research*, 7(Oct):2189–2213, 2006.

Allan Sly and Nike Sun. The computational hardness of counting in two-spin models on d-regular graphs. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 361–369. IEEE, 2012.

Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, DTIC Document, 1986.

David Sontag and Dan Roy. Complexity of inference in latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1008–1016, 2011.

David Steurer. *On the complexity of unique games and graph expansion*. PhD thesis, Princeton University, 2010.

Gilbert W Stewart. Matrix perturbation theory. 1990.

GW Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM review*, 19 (4):634–662, 1977.

R. Sundberg. Maximum likelihood from incomplete data via the em algorithm. *Scandinavian Journal of Statistics*, 1: 49–58, 1974.

M. Telgarsky. Dirichlet draws are sparse with high probability. Manuscript, 2013.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. Tree-reweighted belief propagation algorithms and approximate ml estimation by pseudo-moment matching.

Martin J Wainwright, Tommi S Jaakkola, and Alan S Willsky. A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335, 2005.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Yair Weiss. Correctness of local probability propagation in graphical models with loops. *Neural computation*, 12(1): 1–41, 2000.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.

Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. 2003.

Raymond W Yeung. *Information theory and network coding*. Springer Science & Business Media, 2008.

Yuichi Yoshida and Yuan Zhou. Approximation schemes via sherali-adams hierarchy for dense constraint satisfaction problems and assignment problems. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 423–438. ACM, 2014.