Analyzing Tensor Power Method Dynamics: Applications to Learning Overcomplete Latent Variable Models

Anima Anandkumar∗ Rong Ge† Majid Janzamin‡

November 7, 2014

Abstract
In this paper we provide new guarantees for unsupervised learning of overcomplete latent variable models, where the number of hidden components exceeds the dimensionality of the observed data. In particular, we consider multi-view mixture models and spherical Gaussian mixtures with random mean vectors. Given the third order moment tensor, we learn the parameters using tensor power iterations. We prove that our algorithm can learn the model parameters even when the number of hidden components $k$ is significantly larger than the dimension $d$ up to $k = o(d^{1.5})$, and the signal-to-noise ratio we require is significantly lower than previous results. We present a novel analysis of the dynamics of tensor power iterations, and an efficient characterization of the basin of attraction of the desired local optima.

Keywords: Unsupervised learning, latent variable models, overcomplete representation, tensor decomposition, tensor power method.

1 Introduction
Latent variable models (LVMs), such as topic models, hidden Markov models (HMM) and Gaussian mixtures are ubiquitous in machine learning. Learning these models can provide efficient representation of the observed data, which often leads to higher performance in inference tasks such as classification. Such feature learning strategies are crucial to achieving performance gains in domains such as speech and computer vision (Bengio et al., 2012).

Unsupervised learning is the challenging task of estimating these models when there are no labeled training examples. Spectral techniques have recently found extensive success in unsupervised learning of LVMs, such as HMM (Mossel and Roch, 2005; Hsu et al., 2012), topic models (Anandkumar et al., 2013) and Gaussian mixtures (Hsu and Kakade, 2013). Many of these techniques can be viewed as computing a tensor decomposition on the moment tensor (typically third or fourth order). A usual method in tensor decomposition is tensor power iterations, used to find the tensor eigenvectors. In practice, tensor methods have been effective in a number of applications such as blind source separation (Comon, 2002), topic modeling (Zou et al., 2013), and community detection (Huang et al., 2013).

∗University of California, Irvine. Email: a.anandkumar@uci.edu
†Microsoft Research, New England. Email: rongge@microsoft.com
‡University of California, Irvine. Email: mjanzami@uci.edu
Most of the current results on tensor decomposition are however limited in the following sense: either they require non-degenerate or full-rank factor matrices (Anandkumar et al., 2014c), or they involve higher order moments with large computational and sample complexity requirements (De Lathauwer et al., 2007; Goyal et al., 2013; Bhaskara et al., 2013). Overcomplete LVMs do not satisfy the full rank assumption since the number of hidden components exceeds the observed dimensionality. They often have impressive empirical performance (Coates et al., 2011), can provide greater flexibility in modeling, and are more robust to noise (Lewicki and Sejnowski, 2000).

In this paper, we provide guaranteed approaches for learning overcomplete LVMs through tensor power iterations. This involves establishing global convergence to the true model parameters. Since tensor power iteration is non-convex, there may exist many local optima. We need to characterize the basin of attraction for the true parameters, and provide an efficient approach for initialization which is guaranteed to land in the basin in order to establish global convergence.

In the undercomplete regime, where the factor matrices have full rank, the moment tensor can be orthogonalized, and it is shown that the tensor power iteration does not suffer from any spurious local optima (Anandkumar et al., 2014c). However, this is no longer true for overcomplete LVMs, where the moment tensor cannot be orthogonalized. In this case, there can be many spurious local optima, especially when the number of hidden components grows compared to the observed dimensionality. This makes it challenging to establish guaranteed learning of overcomplete LVMs.

Recently, Anandkumar et al. (2014a,b) provide local convergence guarantees for learning overcomplete LVMs using tensor power iterations. They assume random factor matrices, which results in incoherent tensor components that can be viewed as a soft orthogonality constraint. They establish local convergence when the initialization vectors have a constant amount of correlation with the true components. However, this is limiting since it is hard to obtain such initialization vectors in polynomial time: they can only initialize when the model is mildly overcomplete (i.e., the number of hidden components is at most a constant times the observed dimension) or when given access to samples from the mixture model with a very low amount of noise. In this paper, we significantly improve the analysis of tensor power iteration in (Anandkumar et al., 2014a), and show that the initialization vector only requires a mild amount of correlation with the true tensor components (depending on the level of overcompleteness). Such initialization vectors can be obtained as samples from the mixture model, even when there is a large amount of noise.

1.1 Summary of results

We consider overcomplete multi-view mixture models and spherical Gaussian mixtures. The multi-view mixture (also called naive Bayes) model consists of multiple views which are conditionally independent given the hidden categorical variable. We assume that the mean vectors of the different components are random Gaussian vectors. We analyze convergence of the tensor power iterations on the third order moment tensor. Let $a_j \in \mathbb{R}^d, j \in [k]$ denote the mean vectors of the mixture model, and $\eta \in \mathbb{R}^d$ denote the noise vector for a sample. We establish the following main result:

**Theorem 1** (Informal Result). Consider a multi-view mixture model or a spherical Gaussian mixture with $k$ components in $d$ dimensions, where each mean vector is a random Gaussian vector with expected norm 1. When $k = o(d^{1.5})$ and the sample noise has norm bounded by $\sqrt{d}\sigma$ with high probability, where

$$
\sigma = O \left( \frac{d^{1/2} \log d}{\sqrt{\max\{k, d\}}} \right),
$$

1
for some universal constant \( c > 0 \), then the tensor power iteration converges to vectors close to the true mean vectors.

Note that since the norm of the noise is bounded by \( \sqrt{d}\sigma \), the expected correlation between noise and any component is w.h.p. bounded by \( \sigma \). In particular, for mildly overcomplete models, where \( k = ad \) for some constant \( \alpha > 1 \), the noise bound in any direction is required to be \( \sigma = O\left(2^{-\left(\log d\right)^{1-c}}\right) \), which is larger than any polynomial in \( \frac{1}{d} \). In this case, the norm of the noise is almost \( \sqrt{d} \) times larger than the expected norm of the mean vector. In other words, the signal-to-noise ratio (SNR) can be as low as \( \Omega\left(d^{-1/2+\epsilon}\right) \), for any \( \epsilon > 0 \). Thus, we can learn mixture models with a high level of noise. In general, we establish how the required noise level scales with the number of hidden components \( k \), as long as \( k = o\left(d^{1.5}\right) \).

For mixture of Gaussian models, a series of work starting from [Dasgupta (1999)] gives efficient algorithms for the low noise (well-separated) case. We compare our results with the work by [Arora and Kannan (2005)]. Both our method and [Arora and Kannan (2005)] use samples for initialization, but there are also differences: [Arora and Kannan (2005)] use a spectral-clustering approach and impose separation condition on the mean vectors, while we use a spectral approach based on tensor decomposition and need to impose (stronger) near-orthogonality condition by assuming random mean vectors. In terms of noise, [Arora and Kannan (2005)] can tolerate noise in any direction of up to \( O(d^{-\frac{1}{4}}) \), assuming a constant separation between any mean vectors, while we can tolerate a noise correlation \( \sigma \) as in (1), which is better when \( k = o(d^{1.5}) \). However, we require the additional assumption that the component means are random Gaussian vectors and cannot handle the case when \( k > d^{1.5} \), while the algorithm in [Arora and Kannan (2005)] works for an arbitrary number of components, as long as they are sufficiently separated. On the other hand, our tensor method can handle multi-view mixture models with any arbitrary non-Gaussian noise, as long as it satisfies the noise bound in (1) with high probability. In addition, [Vempala and Wang (2002)] provide a spectral-clustering algorithm for learning spherical Gaussian mixtures where under higher computational complexity as polynomial in \( k^{O(k)} \) and \( d \), their proposed algorithm can tolerate up to constant noise in any direction assuming constant separation between any mean vectors.

**Overview of techniques:** Our analysis of the tensor power iteration relies on the fact that the components of the tensor (which turn out to be the mean vectors of the mixture model) are random Gaussian vectors. Under this assumption, it is not hard to show (see Section 3.1) that the first iteration of tensor power update makes progress. However, after the first iteration, the initial vector and the tensor components are no longer independent of each other. Therefore we cannot directly repeat the same argument for the second step.

How do we analyze the second step even though the vector and tensor components are correlated? The main intuition is to characterize the dependency between the vector and the tensor components, and show that there is still enough randomness left for us to repeat the argument. This idea was inspired by the analysis of approximate message passing (AMP) algorithms [Bayati and Montanari (2010)]. However, our analysis here is very different in several key aspects: 1) In approximate message passing, typically the analysis works in the large system limit, where the number of iterations is fixed and the dimension goes to infinity. Here we can handle a superconstant number of iterations \( O(\log \log d) \), even for finite \( d \); 2) Usually \( k \) is assumed to be a constant factor times \( d \) in the AMP-like analysis, while here we allow them to be polynomially
related.

With the above ideas, we establish that in $O(\log \log d)$ iterations, the tensor power iteration results in a vector with a constant amount of correlation with some true component, when initialized with a vector having a certain small amount of correlation (see the following lemma) with the corresponding true tensor component. This result is the core technical analysis of this paper stated in the following lemma which is a restatement of Lemma 4.

**Lemma 2** (Dynamics of tensor power iteration). Consider rank-$k$ tensor $T = \sum_{j \in [k]} a_j \otimes a_j \otimes a_j$ with random Gaussian components $a_j \sim N(0, \frac{1}{d}I_d)$. Let the initial vector $x^{(1)}$ satisfies the correlation bound

$$|\langle x^{(1)}, a_j \rangle| \geq d^\beta \sqrt{k} d,$$

w.r.t. some true component $a_j, j \in [k]$, for some $\beta > (\log d)^{-c}$ for some universal constant $c > 0$. After $N = \Theta(\log \log d)$ iterations, the tensor power iteration $x \leftarrow \frac{T(T(I, x, x))}{\|T(I, x, x)\|} \quad (see \ (2) \ for \ the \ definition \ of \ this \ multilinear \ form)$ outputs a vector having w.h.p. constant correlation with the true component $a_j$ as $|\langle x^{(N+1)}, a_j \rangle| \geq 1 - \gamma$, for any fixed constant $\gamma > 0$.

It is worth mentioning that the statement in the above Lemma is a general analysis for the dynamics of 3rd order tensor power iteration, and it is not specific to the application of learning latent variable models. In the main theorem for learning multiview mixtures model, we show that the above initialization condition is satisfied when initialized with a sample with noise level as in (1).

Now recall that Anandkumar et al. (2014a) provide convergence of tensor power iteration when the initialization vector has a constant amount of correlation with the true component (see Lemma 5 for the details). Combining our analysis as mentioned above, and the guarantees in (Anandkumar et al., 2014a), we prove that the model components can be recovered up to approximation error $\tilde{O}\left(\sqrt{k}d\right)$. This approximation error occurs in the overcomplete regime since the tensor components are not the fixed points of power iteration. Removing this approximation error is an interesting future direction to investigate. Thus, we provide a novel analysis of the tensor power method with mild initialization requirements for convergence to the true model parameters.

### 1.2 Related work

**Tensor decomposition for learning latent variable models:** In the introduction, some related works are mentioned which study the theoretical and practical aspects of spectral techniques for learning latent variable models. Among them, Anandkumar et al. (2014a) provide the analysis of tensor power iteration for learning several latent variable models in the undercomplete regime, Anandkumar et al. (2014a) provide the analysis in the overcomplete regime and Anandkumar et al. (2014a) provide tensor concentration bounds and apply the analysis in (Anandkumar et al., 2014a) to learning LVMs proposing tight sample complexity guarantees.

**Learning mixture of Gaussians:** Here, we provide a subset of related works studying learning mixture of Gaussians which are more comparable with our result. For a more detailed list of these

---

1In the analysis, we assume that all the weights are equal to one which can be generalized to the case when the ratio of maximum and minimum weights are constant.
works, see Anandkumar et al. (2014c); Hsu and Kakade (2013). The problem of learning mixture of Gaussians dates back to the work by Pearson (1895). They propose a moment-based technique that involves solving systems of multivariate polynomials which is in general challenging in both computational and statistical sense. Recently, lots of studies on learning Gaussian mixture models have been done improving both aspects which can be divided to two main classes: distance-based and spectral methods.

Distance-based methods impose separation condition on the mean vectors showing that under enough separation the parameters can be estimated. Among such approaches, we can mention Dasgupta (1999a); Vempala and Wang (2002); Arora and Kannan (2005). As discussed in the summary of results, these results work even if \( k > d^{1.5} \) as long as the separation condition between means is satisfied, but our work can tolerate higher level of noise in the regime of \( k = o(d^{1.5}) \) with polynomial computational complexity. The guarantees in (Vempala and Wang, 2002) also work in the high noise regime but need higher computational complexity as polynomial in \( k^{O(k)} \) and \( d \).

In the spectral approaches, the observed moments are constructed and the spectral decomposition of the observed moments are performed to recover the parameters (Kalai et al., 2010; Anandkumar et al., 2012, 2014b). Kalai et al. (2010) analyze the problem of learning mixture of two general Gaussians and provide algorithm with high order polynomial sample and computational complexity. Note that in general, the complexity of such methods grow exponentially with the number of components without further assumptions (Moitra and Valiant, 2010). Hsu and Kakade (2013) provide a spectral algorithm under non-degeneracy conditions on the mean vectors and providing guarantees with polynomial sample complexity depending on the condition number of the moment matrices. Anandkumar et al. (2014b) perform tensor power iteration on the third order moment tensor to recover the mean vectors in the overcomplete regime as long as \( k = o(d^{1.5}) \), but need very good initialization vector having constant correlation with the true mean vector. Here, we improve the correlation level required for convergence.

1.3 Notation and tensor preliminaries

Let \([k] := \{1, 2, \ldots, k\}\), and \(\|v\|\) denote the \(\ell_2\) norm of vector \(v\). We use \(\tilde{O}\) and \(\tilde{\Omega}\) to hide polylog factors in asymptotic notations \(O\) and \(\Omega\), respectively.

Tensor preliminaries: A real \(p\)-th order tensor \(T \in \otimes^p \mathbb{R}^d\) is a member of the outer product of Euclidean spaces \(\mathbb{R}^d\). The different dimensions of the tensor are referred to as modes. For instance, for a matrix, the first mode refers to columns and the second mode refers to rows. In addition, fibers are higher order analogues of matrix rows and columns. A fiber is obtained by fixing all but one of the indices of the tensor (and is arranged as a column vector). For example, for a third order tensor \(T \in \mathbb{R}^{d \times d \times d}\), the mode-1 fiber is given by \(T(\cdot, j, l)\). Similarly, slices are obtained by fixing all but two of the indices of the tensor. For example, for the third order tensor \(T\), the slices along 3rd mode are given by \(T(\cdot, \cdot, l)\).

We view a tensor \(T \in \mathbb{R}^{d \times d \times d}\) as a multilinear form. In particular, for vectors \(u, v, w \in \mathbb{R}^d\), we have\(\footnote{Compare with the matrix case where for \(M \in \mathbb{R}^{d \times d}\), we have \(M(I, u) = Mu := \sum_{j \in [d]} u_j M(\cdot, j)\).}
\[T(I, v, w) := \sum_{j,l \in [d]} v_j w_l T(\cdot, j, l) \in \mathbb{R}^d, \tag{2}\]
Figure 1: Multiview mixture model.

which is a multilinear combination of the tensor mode-1 fibers. Similarly \( T(u, v, w) \in \mathbb{R} \) is a multilinear combination of the tensor entries, and \( T(I, I, w) \in \mathbb{R}^{d \times d} \) is a linear combination of the tensor slices.

A 3rd order tensor \( T \in \mathbb{R}^{d \times d \times d} \) is said to be rank-1 if it can be written in the form

\[
T = \lambda \cdot a \otimes b \otimes c \iff T(i, j, l) = \lambda \cdot a(i) \cdot b(j) \cdot c(l),
\]

where notation \( \otimes \) represents the outer product and \( a, b, c \in \mathbb{R}^d \) are unit vectors. A tensor \( T \in \mathbb{R}^{d \times d \times d} \) is said to have a CP rank at most \( k \) if it can be written as the sum of \( k \) rank-1 tensors as

\[
T = \sum_{i \in [k]} \lambda_i a_i \otimes b_i \otimes c_i, \quad \lambda_i \in \mathbb{R}, \ a_i, b_i, c_i \in \mathbb{R}^d.
\]

2 Learning multiview mixture model through tensor methods

In this section, we first introduce the multiview mixture model, and then we propose the learning algorithm and the main guarantees for learning the model.

2.1 Multiview mixture model

Consider an exchangeable multiview mixture model with \( k \) components and \( p \geq 3 \) views; see Figure 1. Suppose that hidden variable \( h \) is a discrete categorical random variable taking one of the \( k \) states. It is convenient to represent it by basis vectors such that

\[ h = e_j \in \mathbb{R}^k \quad \text{if and only if} \quad \text{it takes the } j\text{-th state.} \]

Note that \( e_j \in \mathbb{R}^k \) denotes the \( j \)-the basis vector in the \( k \)-dimensional space. The prior probability for each hidden state is also \( \Pr[h = e_j] = \lambda_j, \ j \in [k] \). For simplicity, in this paper we assume all the \( \lambda_j \)’s are the same. However, similar argument works even when the ratio of maximum and minimum prior probabilities \( \lambda_{\max}/\lambda_{\min} \) is bounded by some constant.

The variables (views) \( z_l \in \mathbb{R}^d \) are related to the hidden state through factor matrix \( A \in \mathbb{R}^{d \times k} \) such that

\[ z_l = Ah + \eta_l, \quad l \in [p], \]

where zero-mean noise vectors \( \eta_l \in \mathbb{R}^d \) are independent of each other and the hidden state \( h \). Given this, the variables (views) \( z_l \in \mathbb{R}^d \) are conditionally independent given the latent variable \( h \), and the conditional means are \( \mathbb{E}[z_l|h = e_j] = a_j \), where \( a_j \in \mathbb{R}^d \) denotes the \( j \)-th column of matrix \( A = [a_1 \cdots a_k] \). In addition, the above properties imply that the order of observations \( z_l \) do not matter and the model is exchangeable. The goal of the learning problem is to recover the parameters of the model (factor matrix) \( A \) given observations.
Algorithm 1 Learning multiview mixture model via tensor power iterations

**Require:** 1) Third order moment tensor $T \in \mathbb{R}^{d \times d \times d}$ in (5), 2) $n$ samples of $z_1$ in multiview mixture model as $z_1^{(\tau)}$, $\tau \in [n]$, and 3) number of iterations $N$.

for $\tau = 1$ to $n$ do
    Initialize unit vectors $x^{(1)}_\tau \leftarrow z^{(\tau)}_1 / \|z^{(\tau)}_1\|$. 
    for $t = 1$ to $N$ do
        Tensor power updates (see (2) for the definition of the multilinear form):
        $$x^{(t+1)}_\tau = T(I, x^{(t)}_\tau, x^{(t)}_\tau) / \|T(I, x^{(t)}_\tau, x^{(t)}_\tau)\|,$$
    end for
end for

return the cluster centers of set $\{x^{(N+1)}_\tau : \tau \in [n]\}$ (by Procedure 2) as estimates $x_j$.

For this model, the third order observed moment has the form (Anandkumar et al., 2014c)
$$E[z_1 \otimes z_2 \otimes z_3] = \sum_{j \in [k]} \lambda_j a_j \otimes a_j \otimes a_j.$$  

Hence, given third order observed moment, the unsupervised learning problem (recovering factor matrix $A$) reduces to computing a tensor decomposition as in (5). The columns $a_j$'s are chosen according to a standard multivariate Gaussian distribution with expected squared norm 1. This Gaussian assumption is standard in the context of approximate message passing algorithms, and there are evidences that similar results could hold for more general distributions; see (Bayati and Montanari, 2010) and references there.

### 2.2 Tensor decomposition algorithm

The algorithm for unsupervised learning of multiview mixture model is based on tensor decomposition techniques provided in Algorithm 1. The main step in (6) performs power iteration \(^4\); see (2) for the multilinear form definition. Note that the algorithm can be thought as the rank-1 form of Alternating Least Square (ALS) where for each initialization the goal is to recover one of the components of the true tensor $T$ denoted by $a_j$. After running the algorithm for all different initialization vectors, the clustering process from (Anandkumar et al., 2014a) ensures that the best converged vectors are returned as the estimation of true components $a_j$.

### 2.3 Main Result

We first propose the conditions required for recovery guarantees, and state a brief explanation of them.

---

\(^3\)It is enough to form the third order moment for our learning purpose.

\(^4\)This is the generalization of matrix power iteration to 3rd order tensors.
**Procedure 2 Clustering process** [Anandkumar et al. 2014a]

**Require:** Tensor $T \in \mathbb{R}^{d \times d \times d}$, set $S := \left\{ x^{(N+1)}_\tau : \tau \in [n] \right\}$, parameter $\nu$.

while $S$ is not empty do
  Choose $x \in S$ which maximizes $|T(x, x, x)|$.
  Do $N$ more iterations of power updates in (6) starting from $x$.
  Let the output of iterations denoted by $\hat{x}$ be the center of a cluster.
  Remove all the $x \in S$ with $|\langle x, \hat{x} \rangle| > \nu/2$.
end while
return the cluster centers.

**Conditions for Theorem 3:**

- Rank condition: $k \leq o(d^{1.5})$.
- The columns of $A$ are uniformly i.i.d. drawn from a standard Gaussian distribution with expected square norm 1, i.e., $a_j \sim \mathcal{N}(0, \frac{1}{d} I_d)$, $j \in [k]$.
- The noise vectors $\eta_l, l \in [3]$, are independent of matrix $A$ and each other. In addition, the norm of noise is w.h.p. bounded by $\sqrt{d}\sigma$ and

$$\sigma \leq O \left( \frac{d^{\frac{1}{2}} - \beta}{\sqrt{\max\{k, d\}}} \right),$$

for some $\beta \geq (\log d)^{-c}$ for universal constant $c > 0$.

The rank condition bounds the level of overcompleteness for which the recovery guarantees are satisfied. The random Gaussian assumption on the columns of $A$ are crucial for analyzing the dynamics of tensor power iteration in the algorithm. We use it to argue there exists enough randomness left in the components after conditioning on the previous iterations. The bound on the noise strength is required to make sure the given sample is close enough to the corresponding mean vector. This ensures that the initial vector is inside the basin-of-attraction of the corresponding component, and hence, the convergence to the mean vector can be guaranteed.

Next, we propose the settings of the algorithm. Let

$$\epsilon_R := \tilde{O} \left( \frac{\sqrt{k}}{d} \right),$$

denote the recovery error.

**Settings of Algorithm 1 in Theorem 3:**

- Number of iterations: $N = \Theta \left( \log \left( \frac{1}{\epsilon_R} \right) + \log \log d \right)$.
- Initialization: As described in the algorithm, the initialization of power iteration is performed by samples of $z_1$ in multiview mixture model $(z^{(1)}_1, \tau \in [n])$ as $x^{(1)}_\tau \leftarrow z^{(1)}_1 / \|z^{(1)}_1\|$.

**Theorem 3.** Assume the conditions and settings mentioned above hold. For any component $a_j, j \in [k]$ such that we have at least one sample $z_1 = a_j + \eta_1$ as the initialization vector \footnote{Note that this happens for component $j$ with high probability when the number of samples is proportional to inverse prior probability corresponding to that mixture.} by using the exact
3rd order moment tensor in \((5)\) as input, Algorithm 1 outputs \(x_j\) as the estimate of \(a_j\) satisfying w.h.p. (over the randomness of the components \(a_j\)'s)

\[\|x_j - a_j\| \leq \epsilon_R,\]

where \(\epsilon_R\) is defined in \((8)\).

Note that the final approximation error in recovery as \(\epsilon_R = \tilde{O}\left(\sqrt{d/k}\right)\) arises due to the fact that the random components of input tensor \(T\) in \((5)\) are not orthogonal to each other, and therefore these components are not fixed points of the power iteration update. In addition, the current algorithm requires the exact third order moment as the input. We conjecture that our algorithm still works even when the algorithm uses an estimation that is inverse polynomially close, but this is left as an open problem.

**Remark 1** (Level of Noise). The expected norm of components \(a_j, j \in [k]\) which are the main signals to be recovered are equal to one. In the regime of constant level of overcompleteness as \(k = \alpha d\) for constant \(\alpha > 1\), the \(\sigma\) is bounded as \(\sigma \leq O\left(d^{-\beta}\right)\) which gives SNR as small as \(d^{-1/2+\epsilon}\) for any \(\epsilon > 0\). Note that Signal-to-noise ratio denoted by SNR is the ratio of signal norm to noise norm. On the other extreme, when \(k = d^{1.5}/(\text{polylog } d)\), the \(\sigma\) is bounded as \(\sigma \leq O\left(d^{-1/4-\beta}\right)\) which gives SNR \(d^{-1/4+\beta}\). This is much better than the previous work by [Anandkumar et al. (2014a)], which only handle noise of norm \(O(1)\), and thus \(\text{SNR} \geq \Omega(1)\).

**Remark 2** (General analysis of third order tensor power iteration). The above theorem and remarks are specific to learning multiview mixture model, while the analysis of power iteration in this paper is more general as long as the initialization condition is satisfied, i.e., when the initial correlation condition in \((10)\) as \(|\langle x^{(1)}, a_j\rangle| \geq d^{\beta} \frac{\sqrt{d}}{\epsilon}\) (for some \(\beta \geq (\log d)^{-c}\)), and other relevant conditions are satisfied, the convergence properties of third order tensor power iteration in Theorem 3 hold.

**Learning spherical Gaussian mixtures:** Consider a mixture of \(k\) different Gaussian vectors with spherical covariance. Let \(a_j \in j \in [k]\) denote the mean vectors and the covariance matrices are \(\sigma^2 I\). Assuming the parameter \(\sigma\) is known, the modified third order observed moment

\[M_3 := \mathbb{E}[z \otimes z \otimes z] - \sigma^2 \sum_{i \in [d]} (\mathbb{E}[z] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[z] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[z])\]

has the tensor decomposition form \(\text{[Hsu and Kakade, 2013]}\)

\[M_3 = \sum_{j \in [k]} \lambda_j a_j \otimes a_j \otimes a_j,\]

where \(\lambda_j\) is the probability of drawing \(j\)-th Gaussian mixture. The above guarantees can be applied to learning mean vectors \(a_j\) in this model with the additional property that the noise is spherical Gaussian. Recall that a comparison on noise tolerance between our method and the work by [Arora and Kannan (2005)], which provide state-of-art results for learning Gaussian mixtures is stated in the summary of results.
Learning multiview mixture model with distinct factor matrices: Consider the multiview mixture model with different factor matrices where the first three views are related to the hidden state as

\[ z_1 = Ah + \eta_1, \quad z_2 = Bh + \eta_2, \quad z_3 = Ch + \eta_3. \]

Then, the guarantees in the above theorem can be extended to recovering the columns of all three factor matrices \( A, B, \) and \( C \) with appropriate modifications in the power iteration algorithm as follows. First the update formula (6) is changed as

\[
\begin{align*}
  x_{1,\tau}^{(t+1)} &= \frac{T \left( I, x_{2,\tau}^{(t)}, x_{3,\tau}^{(t)} \right)}{\left\| T \left( I, x_{2,\tau}^{(t)}, x_{3,\tau}^{(t)} \right) \right\|}, \\
  x_{2,\tau}^{(t+1)} &= \frac{T \left( x_{1,\tau}^{(t)}, I, x_{3,\tau}^{(t)} \right)}{\left\| T \left( x_{1,\tau}^{(t)}, I, x_{3,\tau}^{(t)} \right) \right\|}, \\
  x_{3,\tau}^{(t+1)} &= \frac{T \left( x_{1,\tau}^{(t)}, x_{2,\tau}^{(t)}, I \right)}{\left\| T \left( x_{1,\tau}^{(t)}, x_{2,\tau}^{(t)}, I \right) \right\|},
\end{align*}
\]

which is the alternating asymmetric version of symmetric power iteration in (6). Here, we alternate among different modes of the tensor. In addition, the initialization for each mode of the tensor is appropriately performed with the samples corresponding to that mode. Note that the analysis still works in the asymmetric version since there exists even more independence relationships through the iterations of the power update because of introducing new random matrices \( B \) and \( C \).

3 Proof Outline

The proof of Theorem 3 involves two phases. In the first phase, we show that under certain small amount of correlation (see (10)) between the initial vector and the true component, the power iteration in (6) converges to some vector which has constant correlation with the true component. This result is the core technical analysis of this paper which is provided in Lemma 4. In the second phase, we incorporate the result of Anandkumar et al. (2014a) which guarantees the approximate convergence of power iteration given initial vector having constant correlation with the true component. This is stated in Lemma 5.

To simplify the notation, we consider the tensor

\[
T = \sum_{j \in [k]} a_j \otimes a_j \otimes a_j, \quad a_j \sim \mathcal{N}(0, \frac{1}{d}I_d).
\]

Notice that this is exactly proportional to the 3rd order moment tensor of the multiview mixture model in (5).

**Lemma 4** (Dynamics of tensor power iteration, phase 1). Consider the rank-k tensor \( T \) of the form in (9). Let the initial vector \( x^{(1)} \) satisfies the correlation bound

\[
|\langle x^{(1)}, a_j \rangle| \geq d^3 \frac{\sqrt{k}}{d}, \quad \text{for some true component } a_j, j \in [k], \text{ for some } \beta > (\log d)^{-c} \text{ for some universal constant } c > 0.
\]

After \( N = \Theta (\log \log d) \) iterations, the power update in (6) outputs a vector having w.h.p. constant correlation with the true component \( a_j \) as

\[
|\langle x^{(N+1)}, a_j \rangle| \geq 1 - \gamma,
\]

for any fixed constant \( \gamma > 0. \)

\footnote{In the analysis, we assume that all the weights are equal to one which can be generalized to the case when the ratio of maximum and minimum weights are constant.}
As stated in Remark 2, it is worth emphasizing that the result in Lemma 4 is a general analysis for the dynamics of power iteration, and it is not specific to the application of learning multiview mixture model and the corresponding initialization using samples.

Lemma 5 (Dynamics of tensor power iteration, phase 2 [Anandkumar et al., 2014a]). Consider the rank-\(k\) tensor \(T\) of the form in (1) with rank condition \(k \leq o(d^{1.5})\). Let the initial vector \(x^{(1)}\) satisfies the constant correlation bound

\[
|\langle x^{(1)}, a_j \rangle| \geq 1 - \gamma,
\]

w.r.t. some true component \(a_j, j \in [k]\), for some constant \(\gamma > 0\). Then after \(N = \Theta\left(\log \frac{1}{\epsilon_R}\right)\) iterations, the power update in (6) outputs a vector satisfying w.h.p.\(\mathbb{P}\)

\[
\|x^{(N+1)} - a_j\| \leq \epsilon_R,
\]

where \(\epsilon_R\) is defined in (8).

Given these two lemmas, our main theorem directly follows:

Proof of Theorem 3: The result is proved by combining Lemmas 4 and 5. We only need to show the initialization condition in (10) is w.h.p. satisfied. Without loss of generality, let the sample used as initialization is \(z_1 = a_1 + \eta_1\). We have w.h.p. \(\langle a_1, \eta_1 \rangle \leq \sigma \log d\), hence the correlation between the normalized version of \(z_1\) and \(a_1\) is

\[
\frac{|\langle a_1 + \eta_1, a_1 \rangle|}{\|a_1 + \eta_1\|} \geq 1 - \frac{|\langle \eta_1, a_1 \rangle|}{\|a_1\| + \|\eta_1\|} \geq 1 - \frac{\sigma \log d}{1 + \sqrt{d} \sigma} \geq d^2 \beta \frac{\sqrt{k}}{d},
\]

where triangle inequality is used in the first bound and the bound on \(\sigma\) is exploited in the last inequality. This is the desired initialization condition.

Finally, note that the lower bound requirement on \(\beta\) as \(\beta \geq (\log d)^{-c}\) for universal constant \(c > 0\) is argued in Appendix A.3; see Corollary 2. \(\square\)

3.1 Proof outline of Lemma 4

First step: We first intuitively show the first step of the algorithm makes progress. Suppose the tensor is \(T = \sum_{j \in [k]} a_j \otimes a_j \otimes a_j\), and the initial vector \(x\) has correlation \(|\langle x, a_1 \rangle| \geq d^2 \beta \frac{\sqrt{k}}{d}\) with the first component. The result of the first iteration is the normalized version of the following vector:

\[
\tilde{x} = \sum_{j \in [k]} \langle a_j, x \rangle^2 a_j.
\]

Intuitively, this vector should have roughly \(|\langle a_1, \tilde{x} \rangle|^2 = d^2 \frac{k}{d}\) correlation with \(a_1\) (as the other terms are random they don’t contribute much). On the other hand, the norm of this vector is roughly \(O(\sqrt{k}/d)\): this is because \(|\langle a_j, x \rangle|^2 \) for \(j \neq 1\) is roughly \(1/d\), and the sum of \(k\) random vectors with length \(1/d\) will have length roughly \(O(\sqrt{k}/d)\). These arguments can be made precise showing the normalized version \(\tilde{x}/\|\tilde{x}\|\) has correlation \(d^2 \beta \sqrt{k}/d\) with \(a_1\) ensuring progress in the first step.

\[\text{[Anandkumar et al. (2014a)] recover the vector up to sign since they work in the asymmetric case. In symmetric case it is easy to resolve sign ambiguity issue.}\]

\[\text{[The correlation between two unit Gaussian vectors in } d \text{ dimensions is roughly } 1/\sqrt{d}\]

11
Going forward: As we explained, the basic idea behind proving Lemma 4 is to characterize the conditional distribution of random Gaussian tensor components $a_j$’s given previous iterations. In particular, we show that the residual independent randomness left in these conditional distributions is large enough and we can exploit it to obtain tighter concentration bounds throughout the analysis of the iterations. The Gaussian assumption on the components, and small enough number of iterations are crucial in this argument.

Notations: For two vectors $u, v \in \mathbb{R}^k$, the Hadamard product denoted by $\ast$ is defined as the entry-wise multiplication of vectors, i.e., $(u \ast v)_j := u_j v_j$ for $j \in [k]$. For a matrix $A$, let $P_{LA}$ denote the projection operator to the subspace orthogonal to column span of $A$. For a subspace $R$, let $R^\perp$ denote the space orthogonal to it. Therefore, for a subspace $R$, the projection operator on the subspace orthogonal to $R$ is equivalently denoted by $P_{R^\perp}$ or $P_{x^R}$. For a random matrix $D$, let $D\{u = Dv\}$ denote the conditional distribution of $D$ given linear constraints $u = Dv$.

Lemma 4 involves analyzing the dynamics of power iteration in (6) for 3rd order rank-$k$ tensors. For the rank-$k$ tensor in (1), the power iterative form $x \leftarrow \frac{T(I, x, x)}{\|T(I, x, x)\|}$ can be written as

$$
x^{(t+1)} = A \left( A^\top x^{(t)} \right)^2, \quad (11)
$$

where the multilinear form in (2) is used. Here, $A = [a_1 \cdots a_k] \in \mathbb{R}^{d \times k}$ denotes the factor matrix, and for vector $y \in \mathbb{R}^k$, $y^\circ := y \circ y \in \mathbb{R}^k$ represents the element-wise square of entries of $y$.

We consider the case where $a_i \sim \mathcal{N}(0, \frac{1}{d} I)$ are i.i.d. drawn and we analyze the evolution of the dynamics of the power update. As explained earlier, for a given initialization $x^{(1)}$, the update in the first step can be analyzed easily since $A$ is independent of $x^{(1)}$. However, in subsequent steps, the updates $x^{(t)}$ are dependent on $A$, and it is no longer clear how to provide a tight bound on the evolution of $x^{(t)}$. In this work, we provide a careful analysis by controlling the amount of “correlation build-up” by exploiting the structure of Gaussian matrices under linear constraints. This enables us to provide better guarantees for matrix $A$ with Gaussian entries compared to general matrices $A$.

Intermediate update steps and variables: Before we proceed, we need to break down power update in (11) and introduce some intermediate update steps and variables as follows. Recall that $x^{(1)} \in \mathbb{R}^d$ denotes the initialization vector. Without loss of generality, let us analyze the convergence of power update to first component of rank-$k$ tensor $T$ denoted by $a_1$. Hence, let the first entry of $x^{(1)}$ denoted by $x_1^{(1)}$ be the maximum entry (in absolute value) of $x^{(1)}$, i.e., $x_1^{(1)} = \|x^{(1)}\|_{\infty}$. Let $B := [a_2 \ a_3 \ \cdots \ a_k] \in \mathbb{R}^{d \times (k-1)}$, and therefore $A = [a_1 | B]$. We break the power update formula in (11) into a few steps by introducing intermediate variables $y^{(t)} \in \mathbb{R}^k$ and $\tilde{x}^{(t+1)} \in \mathbb{R}^d$ as

$$
y^{(t)} := A^\top x^{(t)}, \quad \tilde{x}^{(t+1)} := A(y^{(t)})^\circ.
$$

Note that $\tilde{x}^{(t+1)}$ is the unnormalized version of $x^{(t+1)} := \tilde{x}^{(t+1)}/\|\tilde{x}^{(t+1)}\|$, i.e., $\tilde{x}^{(t+1)} := T(I, x^{(t)}, x^{(t)})$. Thus, we need to jointly analyze the dynamics of all variables $x^{(t)}$, $y^{(t)}$ and $(y^{(t)})^\circ$. Define

$$
X^{[t]} := \left[ x^{(1)} \ldots x^{(t)} \right], \quad Y^{[t]} := \left[ y^{(1)} \ldots y^{(t)} \right].
$$
Matrix $B$ is randomly drawn with i.i.d. Gaussian entries $B_{ij} \sim \mathcal{N}(0, \frac{1}{d})$. As the iterations proceed, we consider the following conditional distributions

$$B^{(t,1)} := B|\{X^{[t]}, Y^{[t]}\}, \quad B^{(t,2)} := B|\{X^{[t+1]}, Y^{[t]}\}. \quad (12)$$

Thus, $B^{(t,1)}$ is the conditional distribution of $B$ at the middle of $t^{th}$ iteration (before update step $\tilde{x}^{(t+1)} = A(y^{(t)})^2$) and $B^{(t,2)}$ is the conditional distribution at the end of $t^{th}$ iteration (after update step $\tilde{x}^{(t+1)} = A(y^{(t)})^2$). By analyzing the above conditional distributions, we can characterize the left independent randomness in $B$.

### 3.1.1 Conditional Distributions

In order to characterize the conditional distribution of $B$ under evolution of $x^{(t)}$ and $y^{(t)}$ in (12), we exploit the following basic fact (see (Bayati and Montanari, 2010) for proof).

**Lemma 6 (Conditional distribution of Gaussian matrices under a linear constraint).** Consider random matrix $D$ with i.i.d. Gaussian entries $D_{ij} \sim \mathcal{N}(0, \sigma^2)$. Conditioned on $u = Dv$ with known vectors $u$ and $v$, the matrix $D$ is distributed as

$$D|\{u = Dv\} \overset{(d)}{=} \frac{1}{\|v\|^2} uv^\top + \tilde{D} P_{v},$$

where random matrix $\tilde{D}$ is an independent copy of $D$ with i.i.d. Gaussian entries $\tilde{D}_{ij} \sim \mathcal{N}(0, \sigma^2)$, and $P_{v}$ is the projection operator on to the subspace orthogonal to $v$.

We refer to $\tilde{D} P_{v}$ as the residual random matrix since it represents the remaining randomness left after conditioning. It is a random matrix whose rows are independent random vectors that are orthogonal to $v$, and the variance in each direction orthogonal to $v$ is equal to $\sigma^2$.

The above Lemma can be exploited to characterize the conditional distribution of $B$ introduced in (12). However, a naive direct application using the constraint $Y^{[t]} = A^\top X^{[t]}$ is not transparent for analysis. The reason is the evolution of $x^{(t)}$ and $y^{(t)}$ are themselves governed by the conditional distribution of $B$ given previous iterations. Therefore, we need the following recursive version of Lemma 6.

**Corollary 1 (Iterative conditioning).** Consider random matrix $D$ with i.i.d. Gaussian entries $D_{ij} \sim \mathcal{N}(0, \sigma^2)$, and let $F \overset{(d)}{=} P_{C} DP_{R}$ be the random Gaussian matrix whose columns are orthogonal to space $C$ and rows are orthogonal to space $R$. Conditioned on the linear constraint $u = Dv$, the matrix $F$ is distributed as

$$F|\{u = Dv\} \overset{(d)}{=} \frac{1}{\|P_{R}v\|^2} u(P_{R}v)^\top + P_{C} \tilde{D} P_{\{R,v\}},$$

where random matrix $\tilde{D}$ is an independent copy of $D$ with i.i.d. Gaussian entries $\tilde{D}_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Thus, the residual random matrix $P_{C} \tilde{D} P_{\{R,v\}}$ is a random Gaussian matrix whose columns are orthogonal to $C$ and rows are orthogonal to span$\{R, v\}$. The variance in any remaining dimension is equal to $\sigma^2$. 

13
3.1.2 Form of Iterative Updates

Now we exploit the conditional distribution arguments proposed in the previous section to characterize the conditional distribution of $B$ given the update variables $x$ and $y$ up to the current iteration; recall (12) where $B^{(t,1)}$ is the conditional distribution of $B$ at the middle of $t^{th}$ iteration and $B^{(t,2)}$ at the end of $t^{th}$ iteration. Before that, we need to introduce some more intermediate variables.

Intermedita variables: We separate the first entry of $y$ and $(y)_2^2$ from the rest, i.e., we have

$$y_1^{(t)} = a_1^\top x^{(t)}, \quad y_{-1}^{(t)} = B^\top x^{(t)} \sim (B^{(t-1,2)})^\top x^{(t)},$$

where $y_{-1}^{(t)} \in \mathbb{R}^{k-1}$ denotes $y^{(t)} \in \mathbb{R}^k$ with the first entry removed. The update formula for $\tilde{x}^{(t+1)}$ can be also decomposed as

$$\tilde{x}^{(t+1)} = (y_1^{(t)})^2 a_1 + B w^{(t)} \sim (y_1^{(t)})^2 a_1 + B^{(t,1)} w^{(t)},$$

where

$$w^{(t)} := (y_{-1}^{(t)})^2 \in \mathbb{R}^{k-1},$$

is the new intermediate variable in the power iterations. Let $B_{\text{res.}}^{(t,1)}$ and $B_{\text{res.}}^{(t,2)}$ denote the residual random matrices corresponding to $B^{(t,1)}$ and $B^{(t,2)}$ respectively, and

$$u^{(t+1)} := B_{\text{res.}}^{(t,1)} w^{(t)}, \quad v^{(t)} := (B_{\text{res.}}^{(t-1,2)})^\top x^{(t)},$$

where $u^{(t)} \in \mathbb{R}^d$ and $v^{(t)} \in \mathbb{R}^{k-1}$ are respectively the part of $x^{(t)}$ and $y_{-1}^{(t)}$ representing the residual randomness after conditioning on the previous iterations. We also summarize all variables and notations in Table II in the Appendix which can be used as a reference throughout the paper.

Finally we make the following observations.

Lemma 7 (Form of iterative updates). The conditional distribution of $B$ at the middle of $t^{th}$ iteration denoted by $B^{(t,1)}$ satisfies

$$B^{(t,1)} \overset{(d)}{=} \sum_{i \in [t-1]} \frac{u^{(i+1)}}{\|P_{\perp W^{[i-1]}} w^{(i)}\|^2} \frac{P_{\perp X^{[i-1]}} x^{(i)} (v^{(i)})^\top}{\|P_{\perp X^{[i-1]}} x^{(i)}\|^2} + B_{\text{res.}}^{(t,1)}, \quad (13)$$

$$B_{\text{res.}}^{(t,1)} = P_{\perp X^{[t]}} \tilde{B} P_{\perp W^{[t-1]}}, \quad (14)$$

where random matrix $\tilde{B}$ is an independent copy of $B$ with i.i.d. Gaussian entries $\tilde{B}_{ij} \sim \mathcal{N}(0, \frac{1}{d})$.

Similarly, the conditional distribution of $B$ at the end of $t^{th}$ iteration denoted by $B^{(t,2)}$ satisfies

$$B^{(t,2)} \overset{(d)}{=} \sum_{i \in [t]} \frac{u^{(i+1)}}{\|P_{\perp W^{[i-1]}} w^{(i)}\|^2} \frac{P_{\perp X^{[i-1]}} x^{(i)} (v^{(i)})^\top}{\|P_{\perp X^{[i-1]}} x^{(i)}\|^2} + B_{\text{res.}}^{(t,2)}, \quad (15)$$

$$B_{\text{res.}}^{(t,2)} = P_{\perp X^{[t]}} B' P_{\perp W^{[t]}}, \quad (16)$$

where random matrix $B'$ is an independent copy of $B$ with i.i.d. Gaussian entries $B'_{ij} \sim \mathcal{N}(0, \frac{1}{d})$.

The lemma can be directly proved by applying the iterative conditioning argument in Corollary I. See the detailed proof in the appendix.
Figure 2: Flow of the power update algorithm stating intermediate steps. Iteration $t$ for which the inductive step should be argued is also indicated.

### 3.1.3 Analysis of Iterative Updates

Lemma 7 characterizes the conditional distribution of $B$ given the update variables $x$ and $y$ up to the current iteration; see (12) for the definition of conditional forms of $B$ denoted by $B^{(t,1)}$ and $B^{(t,2)}$. Intuitively, when the number of iterations $t \ll d$, then the residual independent randomness left in $B^{(t,1)}$ and $B^{(t,2)}$ (respectively denoted by $B_{\text{res.}}^{(t,1)}$ and $B_{\text{res.}}^{(t,2)}$) characterized in Lemma 7 is large enough and we can exploit it to obtain tighter concentration bounds throughout the analysis of the iterations.

Note that the goal is to show that under $t \ll d$, the iterations $x^{(t)}$ converge to the true component with constant error, i.e., $|\langle x^{(t)}, a_1 \rangle| \geq 1 - \gamma$ for some constant $\gamma > 0$. If this already holds before iteration $t$ we are done, and if it does not hold, next iteration is analyzed to finally achieve the goal. This analysis is done via induction argument. During the iterations, we maintain several invariants to analyze the dynamics of power update. The goal is to ensure progress in each iteration as in (17).

**Induction hypothesis:** The following are assumed at the beginning of the iteration $t$ as induction hypothesis; see Figure 2 for the scope of inductive step.

1. **Length of Projection on $x$:**

\[
\delta_t \leq \|P_{\perp X^{[t-1]}} x^{(t)} \| \leq 1,
\]

where $\delta_t$ is of order $1/\text{polylog} \ d$, and the value of $\delta_t$ only depends on $t$ and $\log d$.

2. **Length of Projection on $w$:**

\[
\frac{\Delta'_t - 1}{d} \leq \|P_{\perp W^{[t-2]}} w^{(t-1)} \| \leq \Delta'_t \frac{\sqrt{k}}{d},
\]

\[
\|P_{\perp W^{[t-2]}} w^{(t-1)} \|_\infty \leq \Delta'_t \frac{1}{d},
\]

where $\delta'_t$ is of order $1/\text{polylog} \ d$ and $\Delta'_t$ is of order $\text{polylog} \ d$. Both $\delta'_t$ and $\Delta'_t$ only depend on $t$ and $\log d$. 

15
3. Progress:

\[ |\langle a_1, x^{(t)} \rangle| \in [\delta^*_t, \Delta^*_t d^{2t-1} \sqrt{\frac{k}{d}}], \quad (17) \]
\[ \langle a_1, P_{\perp x^{(t-1)}} x^{(t)} \rangle \leq \Delta^*_t d^{2t-1} \frac{\sqrt{k}}{d}. \]

4. Norm of \( u,v \):

\[ \frac{\delta_{t-1}}{2} \sqrt{\frac{k}{d}} \leq \|v^{(t-1)}\| \leq 2\sqrt{\frac{k}{d}}, \]
\[ \frac{\delta'_{t-1}}{2} \sqrt{\frac{k}{d}} \leq \|u^{(t)}\| \leq 2\Delta'_{t-1} \frac{\sqrt{k}}{d}. \]

The analysis for basis of induction and inductive step are provided in Appendix A.

4 Conclusion

In this paper, we provide a novel analysis for the dynamics of third order tensor power iteration showing convergence guarantees to vectors having constant correlation with the tensor component. This enables us to prove unsupervised learning of latent variable models in the challenging overcomplete regime where the hidden dimensionality is larger than the observed dimension. The main technical observation is that under random Gaussian tensor components and small number of iterations, the residual randomness in the components (which are involved in the iterative steps) are sufficiently large. This enables us to show progress in the next iteration of the update step. As future work, it is very interesting to generalize this analysis to higher order tensor power iteration, and more generally to other kinds of iterative updates.

Acknowledgements

A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, NSF Award CCF-1219234, ARO YIP Award W911NF-13-1-0084 and ONR Award N00014-14-1-0665. M. Janzamin is supported by NSF Award CCF-1219234.

\[ ^9 \text{Note that although the bounds on } y^{(t)}_1 \text{ are argued at iteration } t, \text{ the bound on the first entry of } y^{(t)} \text{ denoted by } y^{(t)}_1 = \langle a_1, x^{(t)} \rangle \text{ is assumed here in the induction hypothesis at the end of iteration } t - 1. \]
## Appendix

Table 1: Table of parameters and variables. Superscript $(t)$ denotes the variable at $t$-th iteration.

| Variable | Space | Description | Recursion formula |
|----------|-------|-------------|------------------|
| $A$      | $\mathbb{R}^{d \times k}$ | mapping matrix in update formula (11) | n.a. |
| $x^{(t)}$ | $\mathbb{R}^d$ | update variable in (11) | $x^{(t+1)} := A(y^{(t)})^2$ |
| $y^{(t)}$ | $\mathbb{R}^k$ | intermediate variable in update formula (11) | $y^{(t)} := A^\top x^{(t)}$ |
| $\tilde{x}^{(t)}$ | $\mathbb{R}^d$ | unnormalized version of $x^{(t)}$ | $\tilde{x}^{(t+1)} := A(y^{(t)})^2$ |
| $B$      | $\mathbb{R}^{d \times (k-1)}$ | matrix $A := [a_1 \ a_2 \ \cdots \ a_k]$ with first column removed, i.e., $B := [a_2 \ a_3 \ \cdots \ a_k]$. Note that the first column $a_1$ is the desired one to recover. | n.a. |
| $B^{(t,1)}$ | $\mathbb{R}^{d \times (k-1)}$ | conditional distribution of $B$ given previous iterations at the middle of $t$th iteration (before update step $\tilde{x}^{(t+1)} = A(y^{(t)})^2$) | $B^{(t,1)} := B|\{X^{[t]}, Y^{[t]}\}$ |
| $B^{(t,2)}$ | $\mathbb{R}^{d \times (k-1)}$ | conditional distribution of $B$ given previous iterations at the end of $t$th iteration (after update step $\tilde{x}^{(t+1)} = A(y^{(t)})^2$) | $B^{(t,2)} := B|\{X^{[t+1]}, Y^{[t]}\}$ |
| $B^{\text{res.}}_{(t,1)}$ | $\mathbb{R}^{d \times (k-1)}$ | residual independent randomness left in $B^{(t,1)}$; see Lemma [7] | see equation (14) |
| $B^{\text{res.}}_{(t,2)}$ | $\mathbb{R}^{d \times (k-1)}$ | residual independent randomness left in $B^{(t,2)}$; see Lemma [7] | see equation (16) |
| $w^{(t)}$ | $\mathbb{R}^{k-1}$ | intermediate variable in update formula (11) | $w^{(t)} := (y^{(t)-1})^2$ |
| $u^{(t)}$ | $\mathbb{R}^d$ | part of $x^{(t)}$ representing the left independent randomness | $u^{(t+1)} := B^{\text{res.}}_{(t,1)} w^{(t)}$ |
| $v^{(t)}$ | $\mathbb{R}^{k-1}$ | part of $y^{(t)-1}$ representing the left independent randomness | $v^{(t)} := (B^{\text{res.}}_{(t,2)} \top x^{(t)})$ |

**Proof of Lemma [7]** Recall that we have updates of the form

$$
\tilde{x}^{(t+1)} = A(y^{(t)})^2, \quad w^{(t)} := (y^{(t)-1})^2, \quad y^{(t)} = A^\top x^{(t)}.
$$

Let

$$
X^{[t]} := \left[ x^{(2)} | \ldots | x^{(t)} \right],
$$

Proof of Lemma [7]
and let the rows of $Y^t$ are partitioned as the first and the rest of rows as

$$Y^t = \begin{bmatrix} Y_1^t & | & Y_{\geq 1}^t \end{bmatrix}^\top.$$

We now make the following simple observations

$$B(t, 1) (d) = B|\{Y^t = A^\top X^t, \bar{X}^t\}_{1} = A(Y^{t-1})^\ast 2\}$$

$$= B|\{Y^t = B^\top X^t, \bar{X}^t\}_{1} = a_1(Y^{t-1})^\ast 2 + BW^{t-1}\}$$

$$= B|\{v^{(1)} = B^\top x^{(1)}, \ldots, v^{(t)} = (B_{res.})^{(t-1, 2)} x^{(t)},$$

$$u^{(2)} = B_{res.}^{(1, 1)} w^{(1)}, \ldots, u^{(t)} = B_{res.}^{(t-1, 1)} w^{(t-1)}\}$$

where the second equivalence comes from the fact that $B$ is matrix $A$ with first column removed. Now applying Corollary 1, we have the result. The distribution of $B(t, 2)$ follow similarly.

\[\square\]

### A Analysis of Induction Argument

In this section, we analyze the basis of induction and inductive step for the induction argument proposed in Section 3.1.3 for the proof of Lemma 4.

#### A.1 Basis of induction

We first show that the hypothesis holds for initialization vector $x^{(1)}$ as the basis of induction.

**Claim 1** (Basis of induction). *The induction hypothesis is true for $t = 1$.*

**Proof:** Notice that induction hypothesis for $t = 1$ only involves the bounds on $\|x^{(1)}\|$ and $\langle a_1, x^{(1)} \rangle$ as in Hypotheses $1$ and $2$ respectively. These bounds are directly argued by the correlation assumption on the initial vector $x^{(1)}$ stated in (10) where $\delta_1 = \delta_1^\ast = \Delta_1^\ast = 1$. \[\square\]

#### A.2 Inductive step

Assuming the induction hypothesis holds for all the values till the end of iteration $t - 1$ (stated in Section 3.1.3), we analyze the $t$-th iteration of the algorithm, and prove that induction hypothesis also holds for the values at the end of iteration $t$. See Figure 2 where the scope of iteration $t$ and the flow of the algorithm is shown. In the rest of this section, we pursue the flow of the algorithm at iteration $t$ starting from computing $y^{(t)}$ and ending up with computing $x^{(t+1)}$ to prove the desired induction hypothesis at iteration $t$.

**Hypothesis 4**

We start by showing that the induction Hypothesis 4 holds at iteration $t$ using the induction Hypotheses $1$ and $2$ in the previous iteration.
Claim 2. We have

\[ \frac{\delta_t}{2} \sqrt{\frac{k}{d}} \leq \|v^{(t)}\| \leq 2 \sqrt{\frac{k}{d}}, \]

\[ \frac{\delta_t'}{2} \sqrt{\frac{k}{d}} \leq \|u^{(t+1)}\| \leq 2 \Delta_t' \sqrt{\frac{k}{d}}. \]

Proof: Recall that \( v^{(t)} := (B^{(t-1,2)}_{\text{res}})\top x^{(t)} \), and by applying the form of \( B^{(t-1,2)}_{\text{res}} \) in (16), we have

\[ v^{(t)} = P_{\perp W_{\text{res}}^{(t-1)}} B'\top P_{\perp X_{\text{res}}^{(t-1)}} x^{(t)}. \]  

Since random matrix \( B' \in \mathbb{R}^{d \times (k-1)} \) is an independent copy of \( B \) with i.i.d. Gaussian entries \( B'_{ij} \sim \mathcal{N}(0, 1_d) \), we know \( v^{(t)} \) is a random Gaussian vector in the subspace orthogonal to \( W^{[t-1]} \). On the other hand, for any vector \( z \in \mathbb{R}^d \), we have

\[ \mathbb{E} \left[ \|B'\top z\|^2 \right] = z\top \mathbb{E} [B'\top B'] z = \frac{k-1}{d} \|z\|^2, \]

where \( \mathbb{E} [B'\top B'] = \frac{k-1}{d} I \) is exploited. Let \( z = P_{\perp X_{\text{res}}^{(t-1)}} x^{(t)} \). Then, by applying the above equality to the expansion of \( v^{(t)} \) in (18), we have

\[ \mathbb{E} \left[ \|v^{(t)}\|^2 \right] = \frac{k-t}{k-1} \cdot \frac{k-1}{d} \cdot \|P_{\perp X_{\text{res}}^{(t-1)}} x^{(t)}\|^2 = \frac{k-t}{d} \cdot \|P_{\perp X_{\text{res}}^{(t-1)}} x^{(t)}\|^2 \in \left[ \frac{\delta_t^2 k}{d} \left( 1 - \frac{t}{k} \right), \frac{k}{d} \right], \]

where \( \text{dim} (W^{[t-1]}) = t-1 \) is also used in the first step, and the last step is concluded from Hypothesis 1. Finally, by concentration property of random Gaussian vectors, when \( t \ll d \) we have with high probability

\[ \|v^{(t)}\| \in \left[ \frac{\delta_t}{2} \sqrt{\frac{k}{d}}, 2 \sqrt{\frac{k}{d}} \right]. \]

Similarly, for \( u^{(t+1)} := B_{\text{res}}^{(t,1)} w^{(t)} \), and by applying the form of \( B_{\text{res}}^{(t,1)} \) in (14), we have

\[ u^{(t+1)} = \underbrace{P_{\perp X^{[t]}} \tilde{B} P_{\perp W_{\text{res}}^{(t-1)}} w^{(t)}}_{(19)}. \]

Since random matrix \( \tilde{B} \in \mathbb{R}^{d \times (k-1)} \) is an independent copy of \( B \) with i.i.d. Gaussian entries \( \tilde{B}_{ij} \sim \mathcal{N}(0, 1_d) \), we know \( u^{(t+1)} \) is a random Gaussian vector in the subspace orthogonal to \( X^{[t]} \). On the other hand, for any vector \( z \in \mathbb{R}^{k-1} \), we have

\[ \mathbb{E} \left[ \|\tilde{B} z\|^2 \right] = z\top \mathbb{E} [\tilde{B}\top \tilde{B}] z = \|z\|^2, \]

where \( \mathbb{E} [\tilde{B}\top \tilde{B}] = I \) is exploited. Let \( z = P_{\perp W_{\text{res}}^{(t-1)}} w^{(t)} \). Then, by applying the above equality to the expansion of \( u^{(t+1)} \) in (19), we have

\[ \mathbb{E} \left[ \|u^{(t+1)}\|^2 \right] = \frac{d-t}{d} \cdot \|P_{\perp W_{\text{res}}^{(t-1)}} w^{(t)}\|^2 \in \left[ \left( \frac{\delta_t'}{d} \right)^2 \left( 1 - \frac{t}{d} \right), \left( \Delta_t' \right)^2 \frac{k}{d} \right], \]

where \( \text{dim} (X^{[t]}) = t \) is also used in the first step, and the last step is concluded from Hypothesis 2. Finally, by concentration property of random Gaussian vectors, when \( t \ll d \) we have with high probability

\[ \|u^{(t+1)}\| \in \left[ \frac{\delta_t'}{2} \sqrt{\frac{k}{d}}, 2 \Delta_t' \sqrt{\frac{k}{d}} \right]. \]
Lemma 8. Suppose \( t \) iteration probability \( R \) is a random Gaussian vector. Applying above lemma with Hypothesis \( \beta \) we know \( |y_1^t| = \tilde{\Theta}(d^{2d-1} \sqrt{k/d}) \). The other coordinates of \( y^t := A^T x^t \) are \( y_{\perp 1} = B^T x^t \) which conditioning on the previous iterations are equivalent (in distribution) to

\[
y_{\perp 1}^{(d)} = \left( B^{(t-1,2)} \right)^T x^t
\]

\[
= \left( \sum_{i \in [t-1]} \left( \frac{u^{(i+1)}}{\|P_{\perp W[i-1]} u^{(i)}\|^2} + \frac{P_{\perp X[i-1]} x^{(i)} (v^{(i)})^T}{\|P_{\perp X[i-1]} x^{(i)}\|^2} \right) + D^{(t-1,2)} \right)^T x^t
\]

\[
= \sum_{i \in [t-1]} \left( \tilde{\Theta}(\frac{d^2}{k}) P_{\perp W[i-1]} u^{(i)} \langle u^{(i+1)}, x^{(t)} \rangle + \tilde{\Theta}(1) v^{(i)} \langle P_{\perp X[i-1]} x^{(i)}, x^{(t)} \rangle \right) + v^t, \quad (20)
\]

where form of \( B^{(t-1,2)} \) in (15) is used in the second equality. The bounds on the norms come from Hypotheses \( \mathbb{H} \) and \( \mathbb{I} \). The last term is by definition \( v^t := (D_{\text{res.}}^{t-1,2})^T x^t \). Note that differences in polylog factors in the (upper and lower) bounds in Hypotheses \( \mathbb{H} \) and \( \mathbb{I} \) are represented by notation \( \tilde{\Theta}() \).

We will establish subsequently that if \( k > d \), the terms involving \( v^{(i)} \)'s in the above expansion dominate, and the terms involving \( P_{\perp W[i-1]} u^{(i)} \)'s have norm of a smaller order; see Claim \( \mathbb{K} \).

Computing \( w^t \): In the next step of the algorithm at iteration \( t \), \( w^t \) is computed for which we now argue if the induction hypothesis holds up to iteration \( t \), both lower and upper bounds at iteration \( t \) as \( \|P_{\perp W[i-1]} w^{(t)}\| \in [\delta^t_1, \Delta^t_1 \frac{\sqrt{k}}{d}] \) (see induction Hypothesis \( \mathbb{J} \)) also hold.

Lower bound: For the lower bound, intuitively the fresh random vector \( v^t \) should bring enough randomness into \( w^t \). We formulate that in the following lemma.

Lemma 8. Suppose \( R \) and \( R' \) are two subspaces in \( \mathbb{R}^k \) with dimension at most \( t \leq \frac{k}{16(\log k)^2} \). Let \( p \in \mathbb{R}^k \) be an arbitrary vector, \( z \in \mathbb{R}^k \) be a uniformly random Gaussian vector in the space orthogonal to \( R \), and finally \( w = (p + z) * (p + z) \). Then with high probability, we have

\[
\|P_{\perp R} w\| \geq \frac{E[\|z\|^2]}{40\sqrt{k}}.
\]

Recall that \( w^t := y_{\perp 1}^t * y_{\perp 1}^t \), and \( y_{\perp 1}^t \) is expanded in (20) as sum of an arbitrary vector and a random Gaussian vector. Applying above lemma with \( R = R' = \text{span}(W^{[t-1]}) \), we have with high probability

\[
\|P_{\perp W[i-1]} w^t\| \geq \frac{E[\|v^t\|^2]}{40\sqrt{k}} \geq \frac{\delta^t_1}{160} \sqrt{k/d},
\]

where Hypothesis \( \mathbb{K} \) gives lower bound \( \|v^t\| \geq \delta^t_1 / 2 \sqrt{k/d} \) (used in the second inequality). By choosing \( \delta^t_1 = \delta^2_t / 160 \) the lower bound in Hypothesis \( \mathbb{J} \) is proved.
Upper bound: In order to prove the upper bounds in Hypothesis 2 we follow the sequence of arguments below:

Claim 3 \( \|y^{(t)}_{-1}\|_\infty \leq \frac{t^2}{\delta t} \|w^{(t)}\|_\infty \), by Lemma 9 we know \( \|P_{W[t-1]}w^{(t)}\|_\infty \Rightarrow \|P_{W[t-1]}w^{(t)}\| \)

First we prove a bound on the infinity norm of \( y^{(t)}_{-1} \):

Claim 3 (Upper bound on \( \|y^{(t)}_{-1}\|_\infty \)). We have

\[
\|y^{(t)}_{-1}\|_\infty \leq \frac{t \log d}{\delta t} + (t-1) \left( \frac{\Delta_{t-1}'}{\delta_{t-1}'} \right)^2 \frac{1}{\sqrt{d}} = \tilde{O} \left( \frac{1}{\sqrt{d}} \right).
\]

Proof: We exploit the induction hypothesis to bound the \( \ell_\infty \) norm of all the terms in the expansion of \( y^{(t)}_{-1} \) in (20).

For the terms involving \( v^{(i)} \), since they are random Gaussian vectors with expected square norm at most \( k/d \), by Lemma 13 we know \( \|v^{(i)}\|_\infty \leq \frac{\log d}{\sqrt{d}} \) with high probability. In addition, for \( v^{(i)} \), \( i < t \), the coefficient is bounded as

\[
\frac{\langle P_{\perp X[i-1]}x^{(i)}, x^{(i)} \rangle}{\|P_{\perp X[i-1]}x^{(i)}\|^2} \leq \frac{1}{\|P_{W[i-1]}x^{(i)}\|} \leq \frac{1}{\delta t},
\]

where the last step uses Hypothesis 1. Therefore, the total contribution from terms involving \( v^{(i)} \) in \( \|y^{(t)}_{-1}\|_\infty \) is bounded by \( \frac{t \log d}{\delta t} \).

For the terms involving \( P_{W[i-1]}w^{(i)} \), \( i \in [t-1] \), we have from Hypothesis 2 that the \( \ell_\infty \) norm is bounded as \( \|P_{W[i-1]}w^{(i)}\|_\infty \leq \Delta_{t-1}' \). In addition, the corresponding coefficient is bounded by

\[
\frac{\langle u^{(i+1)}, x^{(i)} \rangle}{\|P_{W[i-1]}w^{(i)}\|^2} \leq \frac{\|u^{(i+1)}\| \cdot \|x^{(i)}\|}{\|P_{W[i-1]}w^{(i)}\|^2} \leq \frac{2\Delta_{t-1}' d}{\delta_{t-1}^2 \sqrt{k}}.
\]

Again bounds in Hypotheses 2 and 4 are exploited in the last inequality. Hence, the total contribution from terms involving \( P_{W[i-1]}w^{(i)} \), \( i \in [t-1] \) in \( \|y^{(t)}_{-1}\|_\infty \) is bounded by \( (t-1) \left( \frac{\Delta_{t-1}'}{\delta_{t-1}'} \right)^2 \frac{1}{\sqrt{k}} \).

Combining the above bounds finishes the proof.

Since \( w^{(t)} := y^{(t)}_{-1} \ast y^{(t)}_{-1} \), the above claim immediately implies that

\[
\|w^{(t)}\|_\infty \leq \tilde{O} \left( \frac{1}{d} \right).
\]

Now we have the \( \ell_\infty \) norm on \( w \), however we need to bound the \( \ell_\infty \) norm of the projected vector \( P_{W[t-1]}w^{(t)} \). Intuitively this is clear as the vectors in the space \( W[t-1] \) all have small \( \ell_\infty \) as guaranteed by induction hypothesis. We formalize this intuition using the following lemma.

Lemma 9. Suppose \( R \) is a subspace in \( \mathbb{R}^k \) of dimension \( t' \), such that there is a basis \( \{r_1, \ldots, r_{t'}\} \) with \( \|r_i\|_\infty \leq \frac{\Delta}{\sqrt{k}} \) and \( \|r_i\| = 1 \). Let \( p \in \mathbb{R}^k \) be an arbitrary vector, then

\[
\|P_{\perp R}p\|_\infty \leq \|p\|_\infty + \|p\| \frac{\Delta \sqrt{t'}}{\sqrt{k}}.
\]
Let $R = \text{span}(W^{[t-1]})$. Then the vectors $P_{\perp W^{[i-1]}} w^{(i)}/\|P_{\perp W^{[i-1]}} w^{(i)}\|$, $i \in [t-1]$ form a basis for subspace $R$, and we know from Hypothesis 2 that the $\ell_{\infty}$ norm of each of these basis vectors is bounded by $\Delta := \frac{\Delta_{t-1}}{\delta}$ which is of order polylog $d$. Applying above lemma, we have

$$\|P_{\perp W^{[i-1]}} w^{(i)}\|_{\infty} \leq \|w^{(i)}\| (1 + \Delta \sqrt{t-1}) \leq \frac{\Delta_{t-1}}{\delta},$$

where the last inequality uses bound (23), and appropriate choosing for $\Delta$ which is of order polylog $d$ and only depends on $t$ and $\log d$. This concludes the upper bound on the $\ell_{\infty}$ norm in Hypothesis 2. The upper bound on the $\ell_2$ norm is also immediately argued using this $\ell_{\infty}$ norm bound where an additional $\sqrt{\kappa}$ factor shows up.

**Hypothesis 1**

**Computing $x^{(t+1)}$:** In the next step of iteration $t$, the algorithm computes $x^{(t+1)}$. Conditioning on the previous iterations, the unnormalized version $\tilde{x}^{(t+1)}$ is equivalent (in distribution) to

$$\tilde{x}^{(t+1)} \overset{(d)}{=} B^{(t,1)} w^{(t)} + (y_1^{(t)})^2 a_1$$

$$= \sum_{i \in [t-1]} \frac{u^{(i+1)} (P_{\perp W^{[i-1]}} w^{(i)})^\top w^{(t)}}{\|P_{\perp W^{[i-1]}} w^{(i)}\|^2} + \sum_{i \in [t]} \frac{P_{\perp X^{[i-1]}} x^{(i)} (v^{(i)})^\top w^{(t)}}{\|P_{\perp X^{[i-1]}} x^{(i)}\|^2} + B^{(t,1)}_{\text{res}} w^{(t)} + (y_1^{(t)})^2 a_1$$

$$= \sum_{i \in [t-1]} \tilde{\Theta} \left( \frac{d^2}{k} \right) u^{(i+1)} (P_{\perp W^{[i-1]}} w^{(i)})^\top w^{(t)} + \sum_{i \in [t]} \tilde{\Theta}(1) P_{\perp X^{[i-1]}} x^{(i)} (v^{(i)})^\top w^{(t)} + u^{(t+1)} + (y_1^{(t)})^2 a_1,$$

(24)

where form of $B^{(t,1)}$ in (13) is used in the second equality. The bounds on the norms come from Hypotheses 1 and 2. The last term is the definition of $u^{(t+1)} := B^{(t,1)}_{\text{res}} w^{(t)}$. Note that differences in polylog factors in the (upper and lower) bounds in Hypotheses 1 and 2 are represented by notation $\Theta(\cdot)$.

The goal is to prove Hypothesis 1 holds at $t$-th iteration (which is to show the desired lower and upper bounds on $\|P_{\perp X^{[t]}} x^{(t+1)}\|$ assuming induction hypothesis holds for earlier iterations. Given the normalization $x^{(t+1)} := \tilde{x}^{(t+1)}/\|\tilde{x}^{(t+1)}\|$ in each iteration, we have

$$\|P_{\perp X^{[t]}} x^{(t+1)}\| = \frac{1}{\|\tilde{x}^{(t+1)}\|} \|P_{\perp X^{[t]}} \tilde{x}^{(t+1)}\|.$$

Therefore, we first bound the norm of $\tilde{x}^{(t+1)}$ which turns out to be $\|\tilde{x}^{(t+1)}\| = \tilde{\Theta} \left( \frac{\sqrt{\kappa}}{d} \right)$ as argued in the following.

**Lower bound:** The lower bound on $\|\tilde{x}^{(t+1)}\|$ simply follows from the term $u^{(t+1)}$, which is an independent random Gaussian.

Claim 4. If $t \leq \frac{d_0}{t_0}$, then we have whp

$$\|\tilde{x}^{(t+1)}\| \geq \frac{\delta}{4} \sqrt{\kappa} \frac{d}{d_0}.$$
Proof: We have
\[ \| \tilde{x}^{(t+1)} \| \geq \| \mathcal{P}_{\text{span}(X[t],U[t],a_1)} \tilde{x}^{(t+1)} \| = \| \mathcal{P}_{\text{span}(X[t],U[t],a_1)} u^{(t+1)} \|. \]
Note that the equality is concluded from expansion of \( \tilde{x}^{(t+1)} \) in (24) where all the components of \( \tilde{x}^{(t+1)} \) in the subspace \( \text{span}(X[t],U[t],a_1) \) is represented by \( u^{(t+1)} \). The vector \( \mathcal{P}_{\text{span}(X[t],U[t],a_1)} u^{(t+1)} \) is the projection of a random Gaussian vector \( u^{(t+1)} \) in to a subspace of dimension \( d - o(d) \). Hence it is still a random Gaussian vector with expected square norm larger than \( \frac{\delta^2 k}{2 - d} \). By Lemma [14] with high probability the desired bound holds.

Upper bound: The upper bound is argued in the following claim.

Claim 5. We have either
\[ \langle x^{(t+1)}, a_1 \rangle \geq 1 - \gamma, \]
for some constant \( \gamma > 0 \) or
\[ \| \tilde{x}^{(t+1)} \| \leq \tilde{O} \left( \frac{\sqrt{k}}{d} \right). \]

Proof: Let \( \tilde{x}^{(t+1)} \) in (24) be written as \( \tilde{x}^{(t+1)} = x + (y_1^{(t)})^2 a_1 \) where vector \( z \in \mathbb{R}^d \) represents all the other terms in the expansion. The analysis is done under two cases 1) \( (y_1^{(t)})^2 \geq \frac{2}{\gamma} \| z \| \) and 2) \( (y_1^{(t)})^2 < \frac{2}{\gamma} \| z \| \) for some constant \( \gamma > 0 \). Note that the left hand side is the norm of \( (y_1^{(t)})^2 a_1 \) since \( \| a_1 \| \), and in addition \( (y_1^{(t)})^2 = \langle x^{(t)}, a_1 \rangle^2 \).

Case 1 \( (y_1^{(t)})^2 \geq \frac{2}{\gamma} \| z \| \): For the \( x^{(t+1)} := \tilde{x}^{(t+1)}/\| \tilde{x}^{(t+1)} \| \), we have
\[ \langle x^{(t+1)}, a_1 \rangle = \frac{1}{\| z + (y_1^{(t)})^2 a_1 \|} (z + (y_1^{(t)})^2 a_1, a_1) \]
\[ \geq \frac{1}{\| z \| + (y_1^{(t)})^2} \left[ (y_1^{(t)})^2 - \| z \| \right] \]
\[ \geq \frac{1 - \frac{2}{\gamma}}{1 + \frac{2}{\gamma}} \geq 1 - \gamma, \]
where triangle and Cauchy-Schwartz inequality are used in the first bound, and the second inequality is concluded from assumption \( (y_1^{(t)})^2 \geq \frac{2}{\gamma} \| z \| \).

Case 2 \( (y_1^{(t)})^2 < \frac{2}{\gamma} \| z \| \): We exploit the induction hypothesis to bound the norm of all the terms in the expansion of \( \tilde{x}^{(t+1)} \) in (24).

For the terms involving \( u^{(i+1)} \), \( i \in [t] \), we have \( \| u^{(i+1)} \| \leq 2 \Delta_i^{\frac{\sqrt{k}}{d}} \) from Hypothesis [4] and the argument for \( \| u^{(t+1)} \| \). In addition, for \( u^{(i+1)} \), \( i \in [t-1] \), the coefficient is bounded as
\[ \frac{\| \mathcal{P}_{\text{span}(w^{(i)},w^{(i)})} u^{(i)} \|}{\| \mathcal{P}_{\text{span}(w^{(i)},w^{(i)})} u^{(i)} \|^2} \leq \frac{\| u^{(i)} \|}{\| \mathcal{P}_{\text{span}(w^{(i)},w^{(i)})} u^{(i)} \|^2} \leq \frac{\Delta_i^2}{\delta_i}, \]
where Cauchy-Schwartz inequality is used in the first bound, and the bound in Hypothesis [2] and (23) are exploited in the last inequality. Therefore, the total contribution from terms involving \( u^{(i+1)} \) in \( \| \tilde{x}^{(t+1)} \| \) is bounded by \( \frac{2(t-1) \Delta_i \sqrt{k}}{\delta_i} \).
For the terms involving $P_{x[i-1]}^\perp x^{(i)}$, $i \in [t]$, we have $\|P_{x[i-1]}^\perp x^{(i)}\| \leq 1$, but the coefficient $\langle v^{(i)}, w^{(t)} \rangle$ needs further analysis to be bounded which is done in Lemma \ref{lem:correlation-bound} saying $|\langle v^{(i)}, w^{(t)} \rangle| \leq \tilde{O} \left( \frac{\sqrt{k}}{d} \right)$. This implies that the total contribution from terms involving $P_{x[i-1]}^\perp x^{(i)}$ in $\|\tilde{x}^{(t+1)}\|$ is bounded by $\tilde{O} \left( \frac{\sqrt{k}}{d} \right)$.

Combining the above bounds and considering the assumption that the norm of $(y_1^{(t)})^2 a_1$ in the expansion of $\tilde{x}^{(t+1)}$ is dominated by the norm of other terms argued above, the proof is complete concluding that $\|\tilde{x}^{(t+1)}\| \leq \tilde{O} \left( \frac{\sqrt{k}}{d} \right)$.

\[ \square \]

**Lemma 10.** Under the induction hypothesis (up to update step $\tilde{x}^{(t+1)} := A y^{(t)+2}$ at iteration $t$), we have for $i \in [t]$, \[ |\langle v^{(i)}, w^{(t)} \rangle| \leq O \left( t^3 \frac{(\Delta _{t-1})^4}{(\delta _{t-1})^4 \delta _t^2} (\log d) \frac{\sqrt{k}}{d} \right) = \tilde{O} \left( \frac{\sqrt{k}}{d} \right). \]

Using \eqref{eq:correlation-bound} and the fact that $\|\tilde{x}^{(t+1)}\| = \tilde{\Theta} \left( \frac{\sqrt{k}}{d} \right)$, we have \[ \|P_{x[i]}^\perp x^{(t+1)}\| \geq \tilde{\Theta} \left( \frac{d}{\sqrt{k}} \right) \|P_{\text{span}(X[i], u^{(t)}, a_1)}^\perp u^{(t+1)}\| \geq \frac{\delta _t'}{4}, \]
where the bound $\|P_{\text{span}(X[i], u^{(t)}, a_1)}^\perp u^{(t+1)}\| \geq \frac{\delta _t'}{4} \frac{\sqrt{k}}{d}$ is also used. This finishes the proof that Hypothesis \ref{hyp:induction-hypothesis} holds.

**Hypothesis \ref{hyp:induction-hypothesis}**

Finally we prove Hypothesis \ref{hyp:induction-hypothesis} at iteration $t$ given earlier induction hypothesis. The first part of the hypothesis is proved in the following claim.

**Claim 6.** We have \[ |\langle a_1, x^{(t+1)} \rangle| \in [\delta _{t+1}^*, \Delta _{t+1}^*] d^{32t} \frac{\sqrt{k}}{d}. \]

**Proof:** We first show the correlation bound on the unnormalized version as $\langle a_1, \tilde{x}^{(t+1)} \rangle$. Looking at the expansion of $\tilde{x}^{(t+1)}$ in \eqref{eq:correlation-expansion}, the correlation $\langle a_1, \tilde{x}^{(t+1)} \rangle$ involves three types of terms emerging from $(y_1^{(t)})^2 a_1$, $u^{(t+1)}$ and $P_{x[i-1]}^\perp x^{(i)}$. In the following, we argue the correlation from each of these terms where we observe that the correlation is dominated by the term $(y_1^{(t)})^2 a_1$, and the rest of terms contribute much smaller amount.

For the term $(y_1^{(t)})^2 a_1$, we have \[ \langle a_1, (y_1^{(t)})^2 a_1 \rangle = (y_1^{(t)})^2 \in [(\delta _t^*)^2, (\Delta _t^*)^2] d^{32t} \frac{k}{d^2}, \]
where the last part exploits induction Hypothesis \ref{hyp:induction-hypothesis} in the previous iteration.

For the terms involving $u^{(t+1)}$, these vectors are random Gaussian vectors in a subspace (with dimension $\Omega(d)$), and therefore, we have with high probability \[ \langle a_1, u^{(t+1)} \rangle \leq \mathbb{E}[\|u^{(t+1)}\|] \cdot O \left( \frac{\log d}{\sqrt{d}} \right) \leq \tilde{O} \left( \frac{\sqrt{k}}{d \sqrt{d}} \right) \leq \tilde{O} \left( \frac{k}{d^2} \right), \]

\[ 24.5 \]
where the correlation bound between two independent random Gaussian vectors in $\Omega(d)$-dimension is used in the first inequality. Hypothesis \textsuperscript{3} is exploited in the second inequality, and finally last inequality is from assumption $k > d$. In addition, the coefficient associated with $u^{(t+1)}$ is bounded by $\Delta'_t/\delta'_t$ argued in \textsuperscript{26}. Hence, the total contribution from terms involving $u^{(t+1)}$ is bounded by $\tilde{O}\left(\frac{\sqrt{k}}{d^t}\right)$.

For the terms involving $P_{\bot X^{(i-1)}}x^{(i)}$, by Hypothesis \textsuperscript{3} we have

$$\langle a_1, P_{\bot X^{(i-1)}}x^{(i)} \rangle \leq \Delta_t^* d^{\beta_2^{-1} - 1} \sqrt{k}.$$  

In addition, the associated coefficient is bounded by $\tilde{O}\left(\frac{\sqrt{k}}{d^t}\right)$ from Lemma \textsuperscript{10}. Hence, the total contribution from terms involving $P_{\bot X^{(i-1)}}x^{(i)}$ in $\langle \tilde{x}^{(t+1)}, a_1 \rangle$ is bounded by $\tilde{O}\left(d^{\beta_2 t - 1} \frac{k}{d^t}\right)$.

Combining the above bounds implies

$$|\langle a_1, \tilde{x}^{(t+1)} \rangle| \leq \tilde{O}\left(d^{\beta_2 t} \frac{k}{d^t}\right).$$

Finally, using the bound on the norm of $\tilde{x}^{(t+1)}$ argued as $\|\tilde{x}^{(t+1)}\| = \tilde{\Theta}\left(\frac{\sqrt{k}}{d}\right)$ finishes the proof. □

To prove the last part of Hypothesis \textsuperscript{3} we use the following lemma which is very similar to Lemma \textsuperscript{9}.

**Lemma 11**. Suppose $R$ is a subspace in $\mathbb{R}^d$ of dimension $t'$, such that there is a basis $\{r_1, \ldots, r_{t'}\}$ with $|\langle r_i, a_1 \rangle| \leq \Delta$ and $\|r_i\| = 1$. Let $p \in \mathbb{R}^d$ be an arbitrary vector, then

$$|\langle p_{\bot R}, a_1 \rangle| \leq |\langle p, a_1 \rangle| + \|p\|\Delta\sqrt{t'}.$$  

We apply this lemma with $R = \text{span}(X^{[t]}), and the basis is $P_{\bot X^{[i-1]}}X^{(i)}/\|P_{\bot X^{[i-1]}}X^{(i)}\|$. By induction hypothesis $\Delta$ in the lemma is at most $\Delta_t' d^{\beta_2} \sqrt{k}/d$, let $v = x^{(t+1)}$ then this gives the desired bound.

Let $R = \text{span}(X^{[t]})$. Then the vectors $P_{\bot X^{[i-1]}}X^{(i)}/\|P_{\bot X^{[i-1]}}X^{(i)}\|$, $i \in [t]$ form a basis for subspace $R$, and we know from Hypotheses \textsuperscript{1} and \textsuperscript{3} that the correlation between these basis vectors and $a_1$ is bounded by $\Delta := \Delta_t' d^{\beta_2 t - 1} \sqrt{k}/d$. Applying above lemma, we have

$$|\langle p_{\bot X^{[t]}}, x^{(t+1)} \rangle| \leq |\langle x^{(t+1)}, a_1 \rangle| + \Delta\sqrt{t} \leq \Delta_{t+1}^* d^{\beta_2 t - 1} \frac{\sqrt{k}}{d},$$

where the last inequality uses the first part of Hypothesis \textsuperscript{3} proved earlier in this section. Note that $\Delta_{t+1}^*$ is a new polylog factor here.

**A.3 Growth rate of $\delta_t$, $\delta_t'$, $\Delta_t'$, $\delta_t^*$, $\Delta_t^*$**

We know that if the number of iterations $t$ is a constant, then the $\delta$ and $\Delta$ parameters (i.e., $\delta_t$, $\delta_t'$, $\Delta_t'$, $\delta_t^*$, $\Delta_t^*$) in the induction hypothesis are bounded by polylog factors of $d$. Here, we show

\footnote{For two independent random Gaussian vectors $p, q \in \mathbb{R}^d$, we have with high probability $\langle p, q \rangle \leq E[\|p\|] \cdot E[\|q\|] \cdot O\left(\frac{\log d}{d}\right)$.}
that these parameters can be still bounded even when the number of steps is slightly larger than a constant. Let

$$R_t := \max\{1/\delta_t, \Delta'_t/\delta'_{t-1}, \Delta_t^{*}/\delta^{*}_t\}.$$  

We know $R_1 = 1$, and by the inductive step analysis we have the following polynomial recursion property.

**Claim 7.** $R_{t+1} = \text{poly}(R_t, t, \log d)$.

This claim follows from the proof of inductive step, where in every step the $\delta$ and $\Delta$ parameters are bounded by polynomial functions of previous $\delta$’s ($\Delta$’s), $t$, and $\log d$.

We now solve this recursion as follows.

**Lemma 12.** Suppose $R_{t+1} \leq c_0 R_t^{c_4} t^{c_2} (\log d)^{c_3}$ where $c_0, c_1, c_2, c_3$ are positive constants, and we know $R_1 = 1$. Then

$$R_t \leq (\log d)^{2c_4 t},$$

for some constant $c_4 > 0$ depending on $c_0, c_1, c_2, c_3$.

**Proof:** Without loss of generality assume $c_0 \geq 1$, $c_2 \geq 1$, $c_3 \geq 1$, and $R_1 \geq \log d$. Given these assumptions, we have $R_t \geq \max\{c_0, t, \log d\}$, for $t \geq 1$. Applying this to the assumption $R_{t+1} \leq c_0 R_t^{c_4} t^{c_2} (\log d)^{c_3}$, we have

$$R_{t+1} \leq R_t^{c_1 + c_2 + c_3}. \quad (27)$$

Pick some $q > 0$ such that $R_1 \leq (\log d)^{2q}$, and pick some

$$c_4 \geq \max\{q, \log_2(1 + c_1 + c_2 + c_3)\}.$$

Now we prove the result by the induction argument. Since $c_4 \geq q$, the basis of induction holds for $R_1$. As the inductive step, suppose $R_t \leq (\log d)^{2c_4 t}$. Applying this to $(27)$, we have

$$R_{t+1} \leq (\log d)^{(1+c_1+c_2+c_3)(2c_4 t)} \leq (\log d)^{2c_4(t+1)},$$

where $2^{c_4} \geq (1 + c_1 + c_2 + c_3)$ is used in the last inequality. This finishes the inductive step and the result is proved.

Using the above bound, we show in the following corollary that the $\delta$ and $\Delta$ parameters in the induction hypothesis are bounded by polylog factors of $d$ even if the number of steps $t$ goes up to $c \log \log d$ for small enough constant $c$. In addition, we show that if $\beta \geq (\log d)^{-c_5}$ for some constant $c_5 > 0$, then the power method converges to a point $x^{(t)}$ which is constant close to the true component.

**Corollary 2.** There exists a universal constant $c_5 > 0$ such that if

$$\beta \geq (\log d)^{-c_5},$$

and the initial correlation is lower bounded by $d^\beta \sqrt{\delta}$ (see (10)), then with high probability the power method gets to a point that is constant close to the true component in $\Theta(\log \log d)$ number of steps.

**Proof:** Pick the number of steps to be $t = (\log \log d)/2c_4$, where $c_4$ is the constant in Lemma 12. Then, from Lemma 12 we have

$$R_t \leq (\log d)^{\sqrt{\log d}} \leq o(d),$$

where the last inequality can be shown by taking the log of both sides. This says that the analysis of inductive step still holds after such number of iterations.

Finally, by progress bound in (17), we can see that if $\beta \geq (\log d)^{-c_5}$, then the power method converges to a point $x^{(t)}$ which is constant close to the true component. ∎
## B Auxiliary Lemmas

In this section we prove the lemmas used in arguing inductive step in Appendix A.2.

We first introduce the following lemma proposing a lower bound on the singular value of product of matrices.

**Lemma 13** (Merikoski and Kumar 2004). Let $C$ and $D$ be $k \times k$ matrices. If $1 \leq i \leq k$ and $1 \leq l \leq k - i + 1$, then
\[
\sigma_i(CD) \geq \sigma_{i+l-1}(C) \cdot \sigma_{k-l+1}(D),
\]
where $\sigma_j(C)$ denotes the $j$-th singular value (in decreasing order) of matrix $C$.

### B.1 Properties of random Gaussian vectors

We start with some basic properties of random Gaussian vectors. First as a simple fact, the norm of a random Gaussian vector is concentrated as follows which is proved via simple concentration inequalities.

**Lemma 14.** Let $z \in \mathbb{R}^d$ be a random Gaussian vector with $E[zz^\top] = \frac{1}{d}I$. Then we have with high probability $\frac{1}{2} \leq \|z\| \leq 2$.

Next we show the $\ell_\infty$ norm of a Gaussian vector is small, even if it is projected on some subspace.

**Lemma 15.** Let $R$ be any linear subspace in $\mathbb{R}^d$ and $z \in \mathbb{R}^d$ be a random Gaussian vector with $E[zz^\top] = \frac{1}{d}I$. Then we have with high probability $\|P_{\perp R}z\|_\infty \leq \log d \sqrt{d}$.

**Proof:** Since $P_{\perp R}$ is a projection matrix, in particular the norm of its columns is bounded by 1. Hence, each entry of $P_{\perp R}z$ is a Gaussian random variable with variance bounded by $\frac{1}{d}$ implying that with high probability the absolute value of each coordinate is smaller than $\log d \sqrt{d}$. Finally, the desired $\ell_\infty$ norm bound is argued by applying union bound.

We can also show that most of the entries are of size at least $\frac{1}{\sqrt{d}}$.

**Lemma 16.** Let $R$ be any linear subspace in $\mathbb{R}^d$ with dimension $t \leq \frac{d}{16 \log d}$ and $z \in \mathbb{R}^d$ be a random Gaussian vector with $E[zz^\top] = \frac{1}{d}I$. Then we have with high probability at least $\frac{1}{2}$ of the entries $i \in [d]$ satisfy $|\langle P_{\perp R}z \rangle_i| \geq \frac{1}{\sqrt{d}}$.

**Proof:** Since the entries of $z$ are independent Gaussian random variables with standard deviation $\frac{1}{\sqrt{d}}$, we know with high probability at least $\frac{1}{2}$ of the entries have absolute value larger than $\frac{1}{2\sqrt{d}}$. On the other hand, $P_Rz$ is also a random Gaussian vector with expected square norm bounded by
\[
E[\|P_Rz\|^2] \leq \frac{16}{16(\log d)^2} = \frac{1}{16(\log d)^2},
\]
where the assumption on the dimension of subspace $R$ is used in the inequality. By Lemma 15 we know with high probability entries of $P_Rz$ are bounded by $1/4\sqrt{d}$. Now $P_{\perp R}z = z - P_Rz$ must have at least $1/2$ of the entries with absolute value larger than $1/4\sqrt{d}$.

Using the above lemmas, we can prove Lemma 8.
Lemma 8 (Restated). Suppose $R$ and $R'$ are two subspaces in $\mathbb{R}^k$ with dimension at most $t \leq \frac{k}{16(\log k)^2}$. Let $p \in \mathbb{R}^k$ be an arbitrary vector, $z \in \mathbb{R}^k$ be a uniformly random Gaussian vector in the space orthogonal to $R$, and finally $w = (p + z) \ast (p + z)$. Then with high probability, we have

$$\|P_{\perp R'}w\| \geq \frac{\mathbb{E}[\|z\|^2]}{40\sqrt{k}}.$$  

Proof: Let $z, z'$ be two independent samples of $z$, and $w, w'$ be the corresponding $w$ vectors. We have

$$w - w' = (p + z) \ast (p + z) - (p + z') \ast (p + z') = (2p + z + z') \ast (z - z').$$  

(28)

By properties of Gaussian vectors, $z + z'$, $z - z'$ are two independent random Gaussian vectors in the subspace orthogonal to $R$ each with expected square norm $2\mathbb{E}[\|z\|^2]$. We use $z_1 := z + z'$ and $z_2 := z - z'$ to denote these two random Gaussian vectors.

Next, we show that with high probability

$$\|P_{\perp R'}(w - w')\| \geq \frac{\mathbb{E}[\|z\|^2]}{20\sqrt{k}}.$$  

Note that this implies the result of lemma as follows. Suppose $\|P_{\perp R'}w\| < \frac{1}{40}\mathbb{E}[\|z\|^2]/\sqrt{k}$ with probability $\delta$. Since $w'$ is an independent sample, with probability $\delta^2$ this bound holds for both $w$ and $w'$. When this happens, we have $\|P_{\perp R'}(w - w')\| < \frac{1}{40}\mathbb{E}[\|z\|^2]/\sqrt{k}$ by triangle inequality. Since we showed $\delta^2$ is negligible, $\delta$ is also negligible.

First we sample $z_2$. Let $R'' = \text{span}(R', p \ast z_2)$. Then by expansion of $w - w'$ in (28), we have

$$\|P_{\perp R'}(w - w')\| = \|P_{\perp R'}(2(p \ast z_2) + (z_1 \ast z_2))\| \geq \|P_{\perp R''}(z_1 \ast z_2)\| = \|P_{\perp R''} \text{Diag}(z_2) P_{\perp R} z_1\|,$$  

(29)

where the inequality is concluded by ignoring the component along $p \ast z_2$ direction. The last equality is from$^{11}$ $u \ast v = \text{Diag}(u) \cdot v$ (for two vectors $u$ and $v$), and the assumption that $z_1 = z + z'$ is in the subspace orthogonal to $R$. For the matrix $P_{\perp R''} \text{Diag}(z_2) P_{\perp R}$, we have$^{12}$

$$\sigma_{k/4} \left(P_{\perp R''} \text{Diag}(z_2) P_{\perp R}\right) \geq \sigma_{k/2} \left(\text{Diag}(z_2)\right) \cdot \sigma_{7k/8} \left(P_{\perp R}\right) \cdot \sigma_{7k/8} \left(P_{\perp R'}\right) \geq \frac{\sqrt{\mathbb{E}[\|z\|^2]}}{4\sqrt{k}},$$

where the first inequality is from Lemma$^{13}$, and the last step is argued as follows. By Lemma$^{14}$, with high probability $z_2$ has square norm at least $\mathbb{E}[\|z\|^2]/2 = \mathbb{E}[\|z\|^2]$, and therefore, by Lemma$^{16}$ at least $k/2$ of its entries have absolute value larger than $\frac{1}{4}\sqrt{\mathbb{E}[\|z\|^2]}/\sqrt{k}$. Therefore, we can restrict attention to the space spanned by the $k/4$ top singular vectors. In addition, within this subspace we have with high probability $\|z_1\|^2 \geq \mathbb{E}[\|z\|^2]/8$, and hence,

$$\|P_{\perp R''} \text{Diag}(z_2) P_{\perp R} z_1\| \geq \frac{\mathbb{E}[\|z\|^2]}{20\sqrt{k}},$$

which finishes the proof by applying (29).

---

$^{11}$For vector $u$, $\text{Diag}(u)$ denotes the diagonal matrix with $u$ as its main diagonal.

$^{12}$Recall that $\sigma_i(A)$ denotes the $i$-th singular value (in decreasing order) of matrix $A$.
B.2 Properties of projections

In this part we prove some basic properties of projections. Intuitively, if the whole subspace has small inner-product with some vector, then the projection of an arbitrary vector to the orthogonal subspace should not change the inner-product with that particular vector by too much. This is what we require in Lemma 11.

Lemma 11 (Restated). Suppose $R$ is a subspace in $\mathbb{R}^d$ of dimension $t'$, such that there is a basis $\{r_1, \ldots, r_{t'}\}$ with $|\langle r_i, a_1 \rangle| \leq \Delta$ and $\|r_i\| = 1$. Let $p \in \mathbb{R}^d$ be an arbitrary vector, then

$$|\langle P_{\perp R}p, a_1 \rangle| \leq |\langle p, a_1 \rangle| + \|p\|\Delta\sqrt{t'}.$$  

Proof: We have $P_{\perp R}p = p - \sum_{i=1}^{t'} \langle p, r_i \rangle r_i$, and therefore

$$|\langle P_{\perp R}p, a_1 \rangle| \leq |\langle p, a_1 \rangle| + \sum_{i=1}^{t'} |\langle p, r_i \rangle\langle a_1, r_i \rangle|$$

$$\leq |\langle p, a_1 \rangle| + \Delta \sum_{i=1}^{t'} |\langle p, r_i \rangle|$$

$$\leq |\langle p, a_1 \rangle| + \Delta \sqrt{\sum_{i=1}^{t'} |\langle p, r_i \rangle|^2}$$

$$\leq |\langle p, a_1 \rangle| + \Delta \|p\|\sqrt{t'}.$$  

The first step is triangle inequality and the third is Cauchy-Schwartz. \(\square\)  

Lemma 9 is very similar.

Lemma 9 (Restated). Suppose $R$ is a subspace in $\mathbb{R}^k$ of dimension $t'$, such that there is a basis $\{r_1, \ldots, r_{t'}\}$ with $\|r_i\|_{\infty} \leq \Delta\sqrt{k}$ and $\|r_i\| = 1$. Let $p \in \mathbb{R}^k$ be an arbitrary vector, then

$$\|P_{\perp R}p\|_{\infty} \leq \|p\|_{\infty} + \|p\|\Delta\frac{\sqrt{t'}}{\sqrt{k}}.$$  

This lemma essentially follows from Lemma 11 because $\ell_{\infty}$ norm is just the maximum inner-product to a basis vector. More specifically, the above lemma is applied for all $a_1 = e_j, j \in [k]$, where $e_j$ denotes the $j$-th basis vector in $\mathbb{R}^k$.

B.3 Bounding correlation between $v$ and $w$

We are only left with Lemma 10. The main difficulty in proving this lemma is that the later steps are dependent on the previous steps. In the proof we show the dependency is bounded and in fact we can treat them as independent.

Lemma 10 (Restated). Under the induction hypothesis (up to update step $\tilde{x}^{(t+1)} := A(y^{(t)})^2$ at iteration $t$), we have for $i \in [t]$,

$$|\langle v^{(i)}, w^{(t)} \rangle| \leq O\left(t^3 \left(\frac{(\Delta'_{t-1})^4}{(\delta'_{t-1})^4(\delta_{t-1})^2} \log d \frac{\sqrt{k}}{d}\right)\right) = \tilde{O}\left(\frac{\sqrt{k}}{d}\right).$$  

29
Proof: Recall $w(t) = y_{-1}^{(t)} * y_{-1}^{(t)}$, and $y_{-1}^{(t)}$ is specified in (21). We now expand the Hadamard product in $w(t)$ and bound all the resulting $O(t^2)$ terms.

The first type of terms has the form $\langle v^{(i)}, P_{\perp W[i_1-i]} w^{(i_1)} * P_{\perp W[i_2-i]} w^{(i_2)} \rangle$, which can be bounded as

$$
\langle v^{(i)}, P_{\perp W[i_1-i]} w^{(i_1)} * P_{\perp W[i_2-i]} w^{(i_2)} \rangle \leq k \cdot \|v^{(i)}\|_\infty \cdot \|P_{\perp W[i_1-i]} w^{(i_1)} * P_{\perp W[i_2-i]} w^{(i_2)}\|_\infty
$$

$$
\leq 2k \log d \frac{(\Delta_{t-1})^2}{d^2},
$$

where $\|v^{(i)}\|_\infty$ is bounded by Lemma 15 and $\ell_\infty$ norm of other vector is bounded by induction Hypothesis 2. In addition, the corresponding coefficient is bounded by (see (22), and note that both $i_1, i_2 < t$)

$$
\frac{4(\Delta_{t-1})^2}{(\delta_{t-1})^4} \frac{d^2}{k}
$$

Hence, the total contribution from such terms is bounded by

$$
8t^2 \frac{(\Delta_{t-1})^2}{(\delta_{t-1})^4} \frac{d^2}{\sqrt{d}}.
$$

(30a)

The second type of terms has the form $\langle v^{(i)}, P_{\perp W[i_1-i]} w^{(i_1)} * v^{(i_2)} \rangle = \langle v^{(i)} * v^{(i_2)}, P_{\perp W[i_1-i]} w^{(i_1)} \rangle$, which can be bounded as

$$
\|P_{\perp W[i_1-i]} w^{(i_1)}\|_\infty \cdot \|v^{(i)} * v^{(i_2)}\|_1 \leq \|P_{\perp W[i_1-i]} w^{(i_1)}\|_\infty \cdot \frac{\|v^{(i)}\|^2 + \|v^{(i_2)}\|^2}{2} \leq 4\Delta_{t-1} \frac{k}{d^2},
$$

where the last inequality is concluded from Hypotheses 2 and 4. In addition, the corresponding coefficient is bounded by (see (21) and (22), and note that both $i_1, i_2 < t$)

$$
\frac{2\Delta_{t-1}}{(\delta_{t-1})^2 \delta_{t-1}} \frac{d}{\sqrt{k}}.
$$

Hence, the total contribution from such terms is bounded by

$$
8t^2 \frac{(\Delta_{t-1})^2}{\delta_{t-1} (\delta_{t-1})^2} \frac{\sqrt{k}}{d^2}.
$$

(30b)

The third type of terms has the form $\langle v^{(i)}, v^{(i_1)} * v^{(i_2)} \rangle$, with coefficient bounded by $1/\delta_{t-1}^2$ (see (21)). For bounding these inner products, we need to use the fact that they are random Gaussian vectors, however the main difficulty is that they are correlated (if $i > j$, then the subspace that $v^{(i)}$ is in that depends on $v^{(j)}$). To resolve this difficulty, we treat $v^{(i)} \in \mathbb{R}^{k-1}$ as projection of $n^{(i)} \in \mathbb{R}^{k-1}$ into subspace orthogonal to $W_t$, where $n^{(i)}$'s are independent Gaussian vectors in the full $k-1$ dimensional space. Independent of the ordering of $i, i_1, i_2$, we have with high probability

$$
\langle n^{(i)}, n^{(i_1)} * n^{(i_2)} \rangle \leq O \left( \frac{\sqrt{k}}{d \sqrt{d}} \right),
$$

30
since it is a sum of $k - 1$ independent mean-0 entries each with variance $\frac{1}{d}$. On the other hand, from Hypothesis [3] we have $E[\|v^{(i)}\|^2] \leq 4\frac{k}{d}$, and since vector $n^{(i)} - v^{(i)}$ is in the subspace $W^{[t-1]}$ with dimension $t$, we have

$$E[\|n^{(i)} - v^{(i)}\|^2] \leq O\left(\frac{t}{k}\right) \cdot \frac{4k}{d} = O\left(\frac{t}{d}\right),$$

and therefore, we have with high probability $\|n^{(i)} - v^{(i)}\| \leq O(\sqrt{t/d})$ for all $i \in [t-1]$. Using this, the difference between $\langle n^{(i)}, n^{(i)} \ast n^{(i+1)} \rangle$ and $\langle v^{(i)}, v^{(i)} \ast v^{(i+1)} \rangle$ can be bounded as

$$|\langle n^{(i)}, n^{(i)} \ast n^{(i+1)} \rangle - \langle v^{(i)}, v^{(i)} \ast v^{(i+1)} \rangle| \leq O\left((\log k)^t \frac{\sqrt{k}}{d \sqrt{d}}\right),$$

where the right hand side is the bound on the dominant term in the expansion of difference as

$$\|\langle n^{(i)}, (n^{(i)} - v^{(i)}) \ast (n^{(i+1)} - v^{(i+1)}) \rangle\| \leq \|n^{(i)}\| \cdot \|(n^{(i)} - v^{(i)}) \ast (n^{(i+1)} - v^{(i+1)})\|$$

$$\leq O\left((\log k)^t \frac{\sqrt{k}}{d} \right) \cdot O\left(\frac{t}{d}\right)$$

$$= O\left((\log k)^t \frac{\sqrt{k}}{d \sqrt{d}}\right).$$

Here, the first inequality is the Cauchy-Schwartz, and the second inequality is from bound on the norm of random Gaussian vector $n^{(i)}$, and the bound on the norm of difference vectors $n^{(i)} - v^{(i)}$ stated earlier. Hence, the total contribution from such terms is bounded by

$$O\left(t^3 \frac{\log k}{d^2} \frac{\sqrt{k}}{d \sqrt{d}}\right). \tag{30c}$$

Taking the sum of all the terms in (30a)-(30c) gives the desired bound.

\[\square\]

References

A. Anandkumar, D. Hsu, and S. M. Kakade. A Method of Moments for Mixture Models and Hidden Markov Models. In Proc. of Conf. on Learning Theory, June 2012.

A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. K. Liu. Two SVDs Suffice: Spectral Decompositions for Probabilistic Topic Modeling and Latent Dirichlet Allocation. to appear in the special issue of Algorithmica on New Theoretical Challenges in Machine Learning, July 2013.

Anima Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. arXiv preprint arXiv:1402.5180, Feb. 2014a.

Anima Anandkumar, Rong Ge, and Majid Janzamin. Provable Learning of Overcomplete Latent Variable Models: Semi-supervised and Unsupervised Settings. arXiv preprint arXiv:1408.0553, Aug. 2014b.
Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014c. URL https://jmlr.org/papers/v15/anandkumar14b.html.

Sanjeev Arora and Ravi Kannan. Learning mixtures of separated nonspherical gaussians. *The Annals of Applied Probability*, 15(1A):69–92, 2005.

Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *arXiv preprint arXiv:1001.3448*, Jan. 2010.

Y. Bengio, A. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538*, 2012.

A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. *arXiv preprint arXiv:1311.3651*, 2013.

Adam Coates, Andrew Y Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

P. Comon. Tensor decompositions. *Mathematics in Signal Processing V*, pages 1–24, 2002.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *FOCS*, 1999a.

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

L. De Lathauwer, J. Castaing, and J.-F. Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *Signal Processing, IEEE Transactions on*, 55(6):2965–2973, 2007.

N. Goyal, S. Vempala, and Y. Xiao. Fourier pca. *arXiv preprint arXiv:1306.5825*, 2013.

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.

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

F. Huang, U. N. Niranjan, M. Hakeem, and A. Anandkumar. Fast Detection of Overlapping Communities via Online Tensor Methods. *ArXiv 1309.0787*, Sept. 2013.

A. T. Kalai, A. Moitra, and G. Valiant. Efficiently learning mixtures of two gaussians. In *STOC*, 2010.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

Jorma K. Merikoski and Ravinder Kumar. Inequalities for spreads of matrix sums and products. *Applied Mathematics E-Notes*, 4:150–159, 2004.
A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of gaussians. In *FOCS*, 2010.

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.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 186:343–414, 1895.

S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *FOCS*, 2002.

J. Y. Zou, D. Hsu, D. C. Parkes, and R. P. Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, pages 2238–2246, 2013.