Spectral Experts for Estimating Mixtures of Linear Regressions

Arun Tejasvi Chaganty CHAGANTY@CS.STANFORD.EDU
Percy Liang PLIANG@CS.STANFORD.EDU
Stanford University, Stanford, CA 94305 USA

Abstract

Discriminative latent-variable models are typically learned using EM or gradient-based optimization, which suffer from local optima. In this paper, we develop a new computationally efficient and provably consistent estimator for a mixture of linear regressions, a simple instance of a discriminative latent-variable model. Our approach relies on a low-rank linear regression to recover a symmetric tensor, which can be factorized into the parameters using a tensor power method. We prove rates of convergence for our estimator and provide an empirical evaluation illustrating its strengths relative to local optimization (EM).

1. Introduction

Discriminative latent-variable models, which combine the high accuracy of discriminative models with the compact expressiveness of latent-variable models, have been widely applied to many tasks, including object recognition (Quattoni et al., 2004), human action recognition (Wang & Mori, 2009), syntactic parsing (Petrov & Klein, 2008), and machine translation (Liang et al., 2006). However, parameter estimation in these models is difficult; past approaches rely on local optimization (EM, gradient descent) and are vulnerable to local optima.

Our broad goal is to develop efficient provably consistent estimators for discriminative latent-variable models. In this paper, we provide a first step in this direction by proposing a new algorithm for a simple model, *a mixture of linear regressions* (Viele & Tong, 2002).

Recently, method of moments estimators have been developed for *generative* latent-variable models, including mixture models, HMMs (Anandkumar et al., 2012b), Latent Dirichlet Allocation (Anandkumar et al., 2012a), and parsing models (Hsu et al., 2012). The basic idea of these methods is to express the unknown model parameters as a tensor factorization of the third-order moments of the model distribution, a quantity which can be estimated from data. The moments have a special symmetric structure which permits the factorization to be computed efficiently using the robust tensor power method (Anandkumar et al., 2012c).

In a mixture of linear regressions, using third-order moments does not directly reveal the tensor structure of the problem, so we cannot simply apply the above tensor factorization techniques. Our approach is to employ low-rank linear regression (Negahban & Wainwright, 2009; Tomioka et al., 2011) to predict the second and third powers of the response. The solution to these regression problems provide the appropriate symmetric tensors, on which we can then apply the tensor power method to retrieve the final parameters.

The result is a simple and efficient two-stage algorithm, which we call Spectral Experts. We prove that our algorithm yields consistent parameter estimates under certain identifiability conditions. We also conduct an empirical evaluation of our technique to understand its statistical properties (Section 5). While Spectral Experts generally does not outperform EM, presumably due to its weaker statistical efficiency, it serves as an effective initialization for EM, significantly outperforming EM with random initialization.

1.1. Notation

Let $[n] = \{1, \dots, n\}$ denote the first n positive integers. We use $O(f(n))$ to denote a function $g(n)$ such that $\lim_{n\to\infty} g(n)/f(n) < \infty$.

We use $x^{\otimes p}$ to represent the p-th order tensor formed by taking the tensor product of $x \in \mathbb{R}^d$; i.e. $x^{\otimes p}_{i_1 \dots i_p} = x_{i_1} \cdots x_{i_p}$. We will use $\langle \cdot, \cdot \rangle$ to denote the generalized dot product between two p-th order tensors:

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

$\langle X, Y \rangle = \sum_{i_1,\dots i_p} X_{i_1,\dots i_p} Y_{i_1,\dots i_p}$. A tensor X is symmetric if for all $i, j \in [d]^p$ which are permutations of each other, $X_{i_1 \cdots i_p} = X_{j_1 \cdots j_p}$ (all tensors in this paper will be symmetric). For a p-th order tensor $X \in (\mathbb{R}^d)^{\otimes p}$, the mode-$i$ unfolding of X is a matrix $X_{(i)} \in \mathbb{R}^{d \times d^{p-1}}$, whose j-th row contains all the elements of X whose i-th index is equal to j.

For a vector X, let $\|X\|_{\text{op}}$ denote the 2-norm. For a matrix X, let $\|X\|_*$ denote the nuclear (trace) norm (sum of singular values), $\|X\|_F$ denote the Frobenius norm (square root of sum of squares of singular values), $\|X\|_{\max}$ denote the max norm (elementwise maximum), $\|X\|_{\text{op}}$ denote the operator norm (largest singular value), and $\sigma_k(X)$ be the k-th largest singular value of X. For a p-th order tensor X, let $\|X\|_* = \frac{1}{p} \sum_{i=1}^p \|X_{(i)}\|_*$ denote the average nuclear norm over all p unfoldings, and let $\|X\|_{\text{op}} = \frac{1}{p} \sum_{i=1}^p \|X_{(i)}\|_{\text{op}}$ denote the average operator norm over all p unfoldings.

For a tensor $X \in (\mathbb{R}^d)^{\otimes p}$, let $\text{cvec}(X) \in \mathbb{R}^{N(d,p)}, N(d,p) = \binom{d+p-1}{p}$ be the collapsed vectorization of X. For example, if $X \in \mathbb{R}^{d \times d}$, $\text{cvec}(X) = (X_{ii} : i \in [d]; X_{ij} + X_{ji} : i, j \in [d], i < j)$. In general, each component of $\text{cvec}(X)$ is indexed by a vector of counts $(c_1, \dots, c_d)$ with total sum $\sum_i c_i = p$. The value of that component is $\sum_{k \in K(c)} X_{k_1 \cdots k_p}$, where $K(c) = \{k \in [d]^p : \forall i \in [d], c_i = |\{j \in [p] : k_j = i\}|\}$ are the set of index vectors k whose count profile is c. For symmetric tensors X and Y, $\langle X, Y \rangle = \langle \text{cvec}(X), \text{cvec}(Y) \rangle$. Later, we'll see that vectorization allow us to perform regression on tensors, and collapsing simplifies our identifiability condition.

2. Model

The mixture of linear regressions model (Viele & Tong, 2002) defines a conditional distribution over a response $y \in \mathbb{R}$ given covariates $x \in \mathbb{R}^d$. Let k be the number of mixture components. The generation of y given x involves three steps: (i) draw a mixture component $h \in [k]$ according to mixture proportions $\pi = (\pi_1, \dots, \pi_k)$; (ii) draw observation noise ϵ from a known zero-mean noise distribution $\mathcal{E}$, and (iii) set y deterministically based on h and ϵ. More compactly:

$$\begin{aligned} h &\sim \text{Multinomial}(\pi), && (1) \\ \epsilon &\sim \mathcal{E}, && (2) \\ y &= \beta_h^T x + \epsilon. && (3) \end{aligned}$$

The parameters of the model are $\theta = (\pi, B)$, where $\pi \in \mathbb{R}^d$ are the mixture proportions and $B = [\beta_1 \mid \cdots \mid \beta_k] \in \mathbb{R}^{d \times k}$ are the regression coefficients. Note that the choice of mixture component h and the observation noise ϵ are independent. The learning problem is stated as follows: given n i.i.d. samples $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$ drawn from the model with some unknown parameters θ^*, return an estimate of the parameters $\hat{\theta}$.

The mixture of linear regressions model has been applied in the statistics literature for modelling music perception, where x is the actual tone and y is the tone perceived by a musician (Viele & Tong, 2002). The model is an instance of the hierarchical mixture of experts (Jacobs et al., 1991), in which the mixture proportions are allowed to depend on x, known as a gating function. This dependence allow the experts to be localized in input space, providing more flexibility, but we do not consider this dependence in our model.

The estimation problem for a mixture of linear regressions is difficult because the mixture components h are unobserved, resulting in a non-convex log marginal likelihood. The parameters are typically learned using expectation maximization (EM) or Gibbs sampling (Viele & Tong, 2002), which suffers from local optima. In the next section, we present a new algorithm that sidesteps the local optima problem entirely.

3. Spectral Experts algorithm

In this section, we describe our Spectral Experts algorithm for estimating model parameters $\theta = (\pi, B)$. The algorithm consists of two steps: (i) low-rank regression to estimate certain symmetric tensors; and (ii) tensor factorization to recover the parameters. The two steps can be performed efficiently using convex optimization and tensor power method, respectively.

To warm up, let us consider linear regression on the response y given x. From the model definition, we have $y = \beta_h^\top x + \epsilon$. The challenge is that the regression coefficients β_h depend on the random h. The first key step is to average over this randomness by defining average regression coefficients $M_1 \stackrel{\text{def}}{=} \sum_{h=1}^k \pi_h \beta_h$. Now we can express y as a linear function of x with non-random coefficients M_1 plus a noise term $\eta_1(x)$:

$$y = \langle M_1, x \rangle + \underbrace{(\langle \beta_h - M_1, x \rangle + \epsilon)}_{\stackrel{\text{def}}{=} \eta_1(x)}. \quad (4)$$

The noise $\eta_1(x)$ is the sum of two terms: (i) the *mixing noise* $\langle M_1 - \beta_h, x \rangle$ due to the random choice of the mixture component h, and (ii) the *observation noise* $\epsilon \sim \mathcal{E}$. Although the noise depends on x, it still has zero mean conditioned on x. We will later show that we can perform linear regression on the data $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ to produce a consistent estimate of M_1. But clearly, knowing M_1 is insufficient for identifying all the parameters θ, as M_1 only contains d degrees of freedom whereas θ contains $O(kd)$.

Intuitively, performing regression on y given x provides only first-order information. The second key insight is that we can perform regression on higher-order powers to obtain more information about the parameters. Specifically, for an integer $p \geq 1$, let us define the average p-th order tensor power of the parameters as follows:

$$M_p \stackrel{\text{def}}{=} \sum_{h=1}^{k} \pi_h \beta_h^{\otimes p}. \tag{5}$$

Now consider performing regression on y^2 given $x^{\otimes 2}$. Expanding $y^2 = (\langle \beta_h, x \rangle + \epsilon)^2$, using the fact that $\langle \beta_h, x \rangle^p = \langle \beta_h^{\otimes p}, x^{\otimes p} \rangle$, we have:

$$\begin{aligned} y^2 &= \langle M_2, x^{\otimes 2} \rangle + \mathbb{E}[\epsilon^2] + \eta_2(x), \qquad (6) \\ \eta_2(x) &= \langle \beta_h^{\otimes 2} - M_2, x^{\otimes 2} \rangle + 2\epsilon \langle \beta_h, x \rangle + (\epsilon^2 - \mathbb{E}[\epsilon^2]). \end{aligned}$$

Again, we have expressed y^2 has a linear function of $x^{\otimes 2}$ with regression coefficients M_2, plus a known bias $\mathbb{E}[\epsilon^2]$ and noise.[1] Importantly, the noise has mean zero; in fact each of the three terms has zero mean by definition of M_2 and independence of ϵ and h.

Performing regression yields a consistent estimate of M_2, but still does not identify all the parameters θ. In particular, B is only identified up to rotation: if $B = [\beta_1 \mid \cdots \mid \beta_k]$ satisfies $B \operatorname{diag}(\pi) B^\top = M_2$ and π is uniform, then $(BQ) \operatorname{diag}(\pi)(Q^\top B^\top) = M_2$ for any orthogonal matrix Q.

Let us now look to the third moment for additional information. We can write y^3 as a linear function of $x^{\otimes 3}$ with coefficients M_3, a known bias $3\,\mathbb{E}[\epsilon^2]\langle \hat{M}_1, x \rangle + \mathbb{E}[\epsilon^3]$ and some noise $\eta_3(x)$:

$$\begin{aligned} y^3 &= \langle M_3, x^{\otimes 3} \rangle + 3\,\mathbb{E}[\epsilon^2]\langle \hat{M}_1, x \rangle + \mathbb{E}[\epsilon^3] + \eta_3(x), \\ \eta_3(x) &= \langle \beta_h^{\otimes 3} - M_3, x^{\otimes 3} \rangle + 3\epsilon \langle \beta_h^{\otimes 2}, x^{\otimes 2} \rangle \qquad (7) \\ &\quad + 3(\epsilon^2 \langle \beta_h, x \rangle - \mathbb{E}[\epsilon^2]\langle \hat{M}_1, x \rangle) + (\epsilon^3 - \mathbb{E}[\epsilon^3]). \end{aligned}$$

The only wrinkle here is that $\eta_3(x)$ does not quite have zero mean. It would if $\hat{M}_1$ were replaced with M_1, but M_1 is not available to us. Nonetheless, as $\hat{M}_1$ concentrates around M_1, the noise bias will go to zero. Performing this regression yields an estimate of M_3. We will see shortly that knowledge of M_2 and M_3 are sufficient to recover all the parameters.

Now we are ready to state our full algorithm, which we call Spectral Experts (Algorithm 1). First, we perform three regressions to recover the *compound parameters* M_1 (4), M_2 (6), and M_3 (7). Since M_2 and M_3 both only have rank k, we can use nuclear norm regularization (Tomioka et al., 2011; Negahban & Wainwright, 2009) to exploit this low-rank structure and improve our compound parameter estimates. In the algorithm, the regularization strengths $\lambda_n^{(2)}$ and $\lambda_n^{(3)}$ are set to $\frac{c}{\sqrt{n}}$ for some constant c.

Having estimated the compound parameters M_1, M_2 and M_3, it remains to recover the original parameters θ. Anandkumar et al. (2012c) showed that for M_2 and M_3 of the forms in (5), it is possible to efficiently accomplish this. Specifically, we first compute a whitening matrix W based on the SVD of M_2 and use that to construct a tensor $T = M_3(W, W, W)$ whose factors are orthogonal. We can use the robust

Algorithm 1 Spectral Experts

input Datasets $\mathcal{D}_p = \{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$ for $p = 1, 2, 3$; regularization strengths $\lambda_n^{(2)}$, $\lambda_n^{(3)}$; observation noise moments $\mathbb{E}[\epsilon^2], \mathbb{E}[\epsilon^3]$.

output Parameters $\hat{\theta} = (\hat{\pi}, [\hat{\beta}_1 \mid \cdots \mid \hat{\beta}_k])$.

1: Estimate compound parameters M_2, M_3 using **low-rank regression**:

$$\hat{M}_1 = \arg\min_{M_1} \quad \frac{1}{2n} \sum_{(x,y) \in \mathcal{D}_1} (\langle M_1, x \rangle - y)^2, \tag{8}$$

$$\hat{M}_2 = \arg\min_{M_2} \quad \lambda_n^{(2)} \|M_2\|_* + \frac{1}{2n} \sum_{(x,y) \in \mathcal{D}_2} (\langle M_2, x^{\otimes 2} \rangle + \mathbb{E}[\epsilon^2] - y^2)^2, \tag{9}$$

$$\hat{M}_3 = \arg\min_{M_3} \quad \lambda_n^{(3)} \|M_3\|_* + \frac{1}{2n} \sum_{(x,y) \in \mathcal{D}_3} (\langle M_3, x^{\otimes 3} \rangle + 3\,\mathbb{E}[\epsilon^2]\langle \hat{M}_1, x \rangle + \mathbb{E}[\epsilon^3] - y^3)^2. \tag{10}$$

2: Estimate parameters $\theta = (\pi, B)$ using **tensor factorization**:

(a) Compute whitening matrix $\hat{W} \in \mathbb{R}^{d \times k}$ (such that $\hat{W}^\top \hat{M}_2 \hat{W} = I$) using SVD.

(b) Compute eigenvalues $\{\hat{a}_h\}_{h=1}^k$ and eigenvectors $\{\hat{v}_h\}_{h=1}^k$ of the whitened tensor $\hat{M}_3(\hat{W}, \hat{W}, \hat{W}) \in \mathbb{R}^{k \times k \times k}$ by using the robust tensor power method.

(c) Return parameter estimates $\hat{\pi}_h = \hat{a}_h^{-2}$ and $\hat{\beta}_h = (\hat{W}^\top)^\dagger (\hat{a}_h \hat{v}_h)$.

[1] If $\mathbb{E}[\epsilon^2]$ were not known, we could treat it as another coefficient to be estimated. The coefficients M_2 and $\mathbb{E}[\epsilon^2]$ can be estimated jointly provided that x does not already contain a bias (x_j must be non-constant for every $j \in [d]$).

tensor power method to compute all the eigenvalues and eigenvectors of T, from which it is easy to recover the parameters π and $\{\beta_h\}$.

Related work In recent years, there has a been a surge of interest in "spectral" methods for learning latent-variable models. One line of work has focused on observable operator models (Hsu et al., 2009; Song et al., 2010; Parikh et al., 2012; Cohen et al., 2012; Balle et al., 2011; Balle & Mohri, 2012) in which a re-parametrization of the true parameters are recovered, which suffices for prediction and density estimation. Another line of work is based on the method of moments and uses eigendecomposition of a certain tensor to recover the parameters (Anandkumar et al., 2012b;a; Hsu et al., 2012; Hsu & Kakade, 2013). Our work extends this second line of work to models that require regression to obtain the desired tensor.

In spirit, Spectral Experts bears some resemblance to the unmixing algorithm for estimation of restricted PCFGs (Hsu et al., 2012). In that work, the observations (moments) provided a linear combination over the compound parameters. "Unmixing" involves solving for the compound parameters by inverting a mixing matrix. In this work, each data point (appropriately transformed) provides a different noisy projection of the compound parameters.

Other work has focused on learning discriminative models, notably Balle et al. (2011) for finite state transducers (functions from strings to strings), and Balle & Mohri (2012) for weighted finite state automata (functions from strings to real numbers). Similar to Spectral Experts, Balle & Mohri (2012) used a two-step approach, where convex optimization is first used to estimate moments (the Hankel matrix in their case), after which these moments are subjected to spectral decomposition. However, these methods are developed in the observable operator framework, whereas we consider parameter estimation.

The idea of performing low-rank regression on y^2 has been explored in the context of signal recovery from magnitude measurements (Candes et al., 2011; Ohlsson et al., 2012). There, the actual observed response was y^2, whereas in our case, we deliberately construct powers y, y^2, y^3 to identify the underlying parameters.

4. Theoretical results

In this section, we provide theoretical guarantees for the Spectral Experts algorithm. Our main result shows that the parameter estimates $\hat{\theta}$ converge to θ at a $\frac{1}{\sqrt{n}}$ rate that depends polynomially on the bounds on the parameters, covariates, and noise, as well the k-th smallest singular values of the compound parameters and various covariance matrices.

Theorem 1 (Convergence of Spectral Experts)**.** *Assume each dataset $\mathcal{D}_p$ (for $p = 1, 2, 3$) consists of n i.i.d. points independently drawn from a mixture of linear regressions model with parameter θ^*.*[2] *Further, assume $\|x\|_2 \le R$, $\|\beta_h^*\|_2 \le L$ for all $h \in [k]$, $|\epsilon| \le S$ and B is rank k. Let $\Sigma_p \stackrel{\text{def}}{=} \mathbb{E}[\text{cvec}(x^{\otimes p})^{\otimes 2}]$, and assume $\Sigma_p \succ 0$ for each $p \in \{1, 2, 3\}$. Let $\epsilon < \frac{1}{2}$. Suppose the number of samples is $n = \max(n_1, n_2)$ where*

$$n_1 = \Omega\left(\frac{R^{12}\log(1/\delta)}{\min_{p\in[3]}\sigma_{\min}(\Sigma_p)^2}\right)$$

$$n_2 = \Omega\left(\epsilon^{-2}\ \frac{k^2\pi_{\max}^2\|M_2\|_{\text{op}}^{1/2}\|M_3\|_{\text{op}}^2 L^6 S^6 R^{12}}{\sigma_k(M_2)^5\sigma_{\min}(\Sigma_1)^2}\log(1/\delta)\right).$$

If each regularization strength $\lambda_n^{(p)}$ is set to

$$\Theta\left(\frac{L^p S^p R^{2p}}{\sigma_{\min}(\Sigma_1)^2}\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

for $p \in 2, 3$, then the parameter estimates $\hat{\theta} = (\hat{\pi}, \hat{B})$ returned by Algorithm 1 (with the columns appropriately permuted) satisfies

$$\|\hat{\pi} - \pi\|_\infty \le \epsilon \qquad \|\hat{\beta}_h - \beta_h\|_2 \le \epsilon$$

for all $h \in [k]$.

While the dependence on some of the norms (L^6, S^6, R^{12}) looks formidable, it is in some sense unavoidable, since we need to perform regression on third-order moments. Classically, the number of samples required is squared norm of the covariance matrix, which itself is bounded by the squared norm of the data, R^3. This third-order dependence also shows up in the regularization strengths; the cubic terms bound each of ϵ^3, β_h^3 and $\|(x^{\otimes 3})^{\otimes 2}\|_F$ with high probability.

The proof of the theorem has two parts. First, we bound the error in the compound parameters estimates $\hat{M}_2, \hat{M}_3$ using results from Tomioka et al. (2011). Then we use results from Anandkumar et al. (2012c) to convert this error into a bound on the actual parameter estimates $\hat{\theta} = (\hat{\pi}, \hat{B})$ derived from the robust tensor power method. But first, let us study a more basic property: identifiability.

4.1. Identifiability from moments

In ordinary linear regression, the regression coefficients $\beta \in \mathbb{R}^d$ are identifiable if and only if the data has

[2]Having three independent copies simplifies the analysis.

full rank: $\mathbb{E}[x^{\otimes 2}] \succ 0$, and furthermore, identifying β requires only moments $\mathbb{E}[xy]$ and $\mathbb{E}[x^{\otimes 2}]$ (by observing the optimality conditions for (4)). However, in mixture of linear regressions, these two moments only allow us to recover M_1. Theorem 1 shows that if we have the higher order analogues, $\mathbb{E}[x^{\otimes p} y^{\otimes p}]$ and $\mathbb{E}[x^{\otimes 2p}]$ for $p \in \{1, 2, 3\}$, we can then identify the parameters $\theta = (\pi, B)$, provided the following *identifiability condition* holds: $\mathbb{E}[\mathrm{cvec}(x^{\otimes p})^{\otimes 2}] \succ 0$ for $p \in \{1, 2, 3\}$.

This identifiability condition warrants a little care, as we can run into trouble when components of x are dependent on each other in a particular algebraic way. For example, suppose $x = (1, t, t^2)$, the common polynomial basis expansion, so that all the coordinates are deterministically related. While $\mathbb{E}[x^{\otimes 2}] \succ 0$ might be satisfied (sufficient for ordinary linear regression), $\mathbb{E}[\mathrm{cvec}(x^{\otimes 2})^{\otimes 2}]$ is singular for any data distribution. To see this, note that $\mathrm{cvec}(x^{\otimes 2}) = [1 \cdot 1, t \cdot t, 2(1 \cdot t^2), 2(t \cdot t^2), (t^2 \cdot t^2)]$ contains components $t \cdot t$ and $2(1 \cdot t^2)$, which are linearly dependent. Therefore, Spectral Experts would not be able to identify the parameters of a mixture of linear regressions for this data distribution.

We can show that some amount of unidentifiability is intrinsic to estimation from low-order moments, not just an artefact of our estimation procedure. Suppose $x = (t, \dots, t^d)$. Even if we observed all moments $\mathbb{E}[x^{\otimes p} y^{\otimes p}]$ and $\mathbb{E}[x^{\otimes 2p}]$ for $p \in [r]$ for some r, all the resulting coordinates would be monomials of t up to only degree $2dr$, and thus the moments live in a $2dr$-dimensional subspace. On the other hand, the parameters θ live in a subspace of at least dimension dk. Therefore, at least $r \geq k/2$ moments are required for identifiability of any algorithm for this monomial example.

4.2. Analysis of low-rank regression

In this section, we will bound the error of the compound parameter estimates $\|\Delta_2\|_F^2$ and $\|\Delta_3\|_F^2$, where $\Delta_2 \stackrel{\text{def}}{=} \hat{M}_2 - M_2$ and $\Delta_3 \stackrel{\text{def}}{=} \hat{M}_3 - M_3$. Our analysis is based on the low-rank regression framework of Tomioka et al. (2011) for tensors, which builds on Negahban & Wainwright (2009) for matrices. The main calculation involved is controlling the noise $\eta_p(x)$, which involves various polynomial combinations of the mixing noise and observation noise.

Let us first establish some notation that unifies the three regressions ((8), (9), and (10)). Define the observation operator $\mathfrak{X}_p(M_p) : \mathbb{R}^{d^{\otimes p}} \to \mathbb{R}^n$ mapping compound parameters M_p:

$$\mathfrak{X}_p(M_p; \mathcal{D})_i \stackrel{\text{def}}{=} \langle M_p, x_i^{\otimes p} \rangle, \qquad (x_i, y_i) \in \mathcal{D}. \tag{11}$$

Let $\kappa(\mathfrak{X}_p)$ be the restricted strong convexity constant, and let $\mathfrak{X}_p^*(\eta_p; \mathcal{D}) = \sum_{(x,y) \in \mathcal{D}} \eta_p(x) x^{\otimes p}$ be the adjoint.

Lemma 1 (Tomioka et al. (2011), Theorem 1)**.** *Suppose there exists a restricted strong convexity constant $\kappa(\mathfrak{X}_p)$ such that*

$$\frac{1}{n} \| \mathfrak{X}_p(\Delta) \|_2^2 \geq \kappa(\mathfrak{X}_p) \|\Delta\|_F^2 \quad \textit{and} \quad \lambda_n^{(p)} \geq \frac{2 \| \mathfrak{X}_p^*(\eta_p) \|_{\mathrm{op}}}{n}.$$

Then the error of $\hat{M}_p$ is bounded as follows:

$$\|\hat{M}_p - M_p\|_F \leq \frac{32 \lambda_n^{(p)} \sqrt{k}}{\kappa(\mathfrak{X}_p)}.$$

Going forward, we need to lower bound the restricted strong convexity constant $\kappa(\mathfrak{X}_p)$ and upper bound the operator norm of the adjoint operator $\| \mathfrak{X}_p^*(\eta_p) \|_{\mathrm{op}}$. The proofs of the following lemmas follow from standard concentration inequalities and are detailed in Appendix A.

Lemma 2 (lower bound on restricted strong convexity constant)**.** *If*

$$n = \Omega\left(\max_{p \in [3]} \frac{R^{4p} (p!)^2 \log(1/\delta)}{\sigma_{\min}(\Sigma_p)^2} \right),$$

then with probability at least $1 - \delta$:

$$\kappa(\mathfrak{X}_p) \geq \frac{\sigma_{\min}(\Sigma_p)}{2},$$

for each $p \in [3]$.

Lemma 3 (upper bound on adjoint operator)**.** *If*

$$n = \Omega\left(\max_{p \in [3]} \frac{L^{2p} S^{2p} R^{4p} \log(1/\delta)}{\sigma_{\min}(\Sigma_1)^2 \left(\lambda_n^{(p)}\right)^2} \right),$$

then with probability at least $1 - \delta$:

$$\lambda_n^{(p)} \geq \frac{1}{n} \| \mathfrak{X}_p^*(\eta_p) \|_{\mathrm{op}},$$

for each $p \in [3]$.

4.3. Analysis of the tensor factorization

Having bounded the error of the compound parameter estimates $\hat{M}_2$ and $\hat{M}_3$, we will now study how this error propagates through the tensor factorization step of Algorithm 1, which includes whitening, applying the robust tensor power method (Anandkumar et al., 2012c), and unwhitening.

Lemma 4. *Let $M_3 = \sum_{h=1}^k \pi_h \beta_h^{\otimes 3}$. Let $\|\hat{M}_2 - M_2\|_{\mathrm{op}}$ and $\|\hat{M}_3 - M_3\|_{\mathrm{op}}$ both be less than*

$$\frac{\sigma_k(M_2)^{5/2}}{k\pi_{\max}\|M_2\|_{\text{op}}^{1/2}\|M_3\|_{\text{op}}}\ \epsilon,$$

for some $\epsilon < \frac{1}{2}$*. Then, there exists a permutation of indices such that the parameter estimates found in step 2 of Algorithm 1 satisfy the following with probability at least* $1 - \delta$*:*

$$\|\hat{\pi} - \pi\|_\infty \leq \epsilon$$
$$\|\hat{\beta}_h - \beta_h\|_2 \leq \epsilon.$$

for all $h \in [k]$*.*

The proof follows by applying standard matrix perturbation results for the whitening and unwhitening operators and can be found in Appendix B.

4.4. Synthesis

Together, these lemmas allow us to control the compound parameter error and the recovery error. We now apply them in the proof of Theorem 1:

Proof of Theorem 1 (sketch). By Lemma 1, Lemma 5 and Lemma 6, we can control the Frobenius norm of the error in the moments, which directly upper bounds the operator norm: If $n \geq \max\{n_1, n_2\}$, then

$$\|\hat{M}_p - M_p\|_{\text{op}} = O\left(\lambda_n^{(p)}\sqrt{k}\sigma_{\min}(\Sigma_p)^{-1}\right). \qquad (12)$$

We complete the proof by applying Lemma 7 with the above bound on $\|\hat{M}_p - M_p\|_{\text{op}}$.

□

5. Empirical evaluation

In the previous section, we showed that Spectral Experts provides a consistent estimator. In this section, we explore the empirical properties of our algorithm on simulated data. Our main finding is that Spectral Experts alone attains higher parameter error than EM, but this is not the complete story. If we initialize EM with the estimates returned by Spectral Experts, then we end up with much better estimates than EM from a random initialization.

5.1. Experimental setup

Algorithms We experimented with three algorithms. The first algorithm (Spectral) is simply the Spectral Experts. We set the regularization strengths $\lambda_n^{(2)} = \frac{1}{10^5\sqrt{n}}$ and $\lambda_n^{(3)} = \frac{1}{10^3\sqrt{n}}$; the algorithm was not very sensitive to these choices. We solved the low-rank regression to estimate M_2 and M_3 using an off-the-shelf convex optimizer, CVX (Grant & Boyd, 2012). The second algorithm (EM) is EM where the β's are initialized from a standard normal and π was set to the uniform distribution plus some small perturbations. We ran EM for 1000 iterations. In the final algorithm (Spectral+EM), we initialized EM with the output of Spectral Experts.

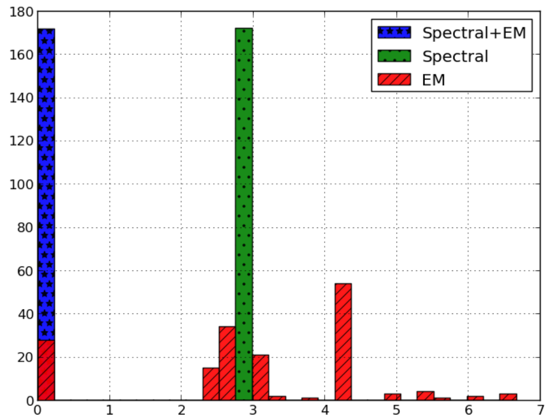


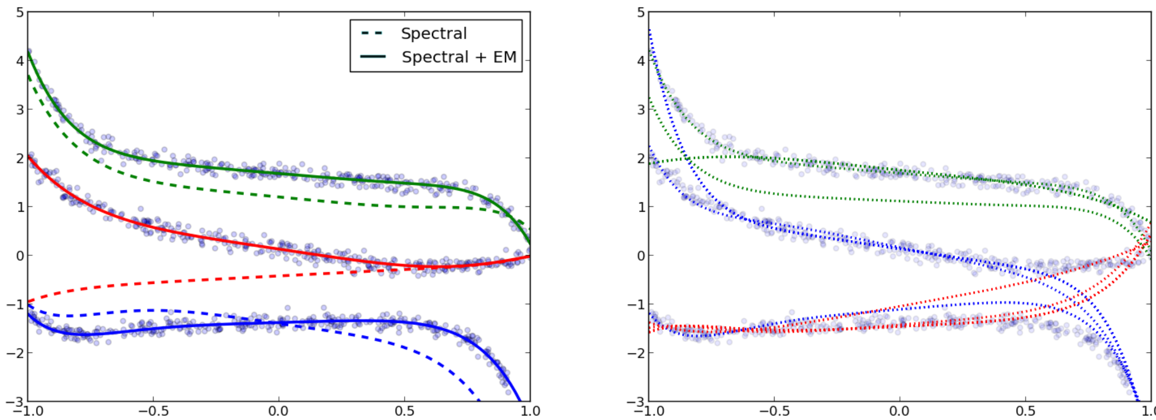
Figure 1. Histogram over recovery errors between the three algorithms when $b = 1, d = 4, k = 3, n = 500,000$.

Data We generated synthetic data as follows: First, we generated a vector t sampled uniformly over the b-dimensional unit hypercube $[-1, 1]^b$. Then, to get the actual covariates x, we applied a non-linear function of t that conformed to the identifiability criteria discussed in Section 3. The true regression coefficients $\{\beta_h\}$ were drawn from a standard normal and π is set to the uniform distribution. The observation noise ϵ is drawn from a normal with variance σ^2. Results are presented below for $\sigma^2 = 0.1$, but we did not observe any qualitatively different behavior for choices of σ^2 in the range $[0.01, 0.4]$.

As an example, one feature map we considered in the one-dimensional setting ($b = 1$) was $x = (1, t, t^4, t^7)$. The data and the curves fit using Spectral Experts, EM with random initialization and EM initialized with the parameters recovered using Spectral Experts are shown in Figure 2. We note that even on well-separated data such as this, EM converged to the correct basin of attraction only 13% of the time.

5.2. Results

Table 1 presents the Frobenius norm of the difference between true and estimated parameters for the model, averaged over 20 different random instances for each



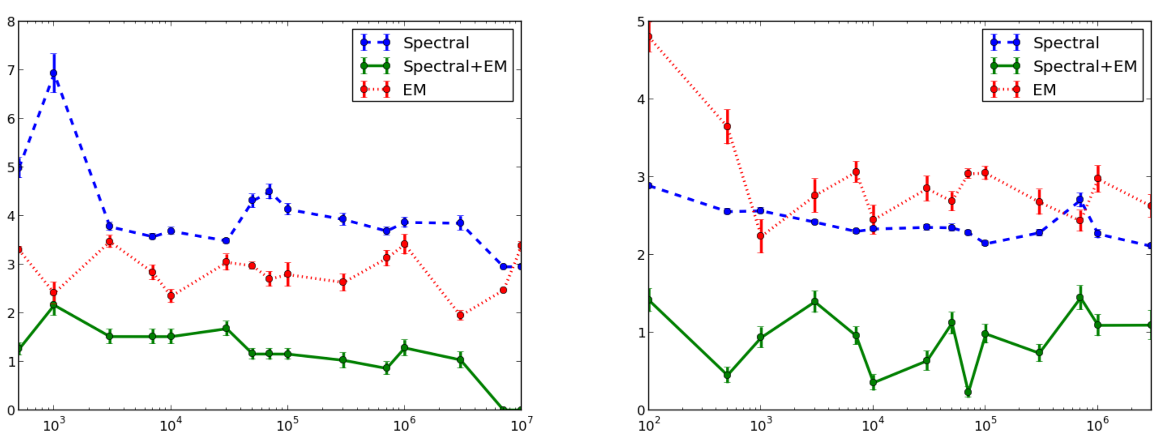
(a) Spectral, Spectral+EM

(b) EM

Figure 2. Visualization of the parameters estimated by Spectral Experts versus EM. (a) The dashed lines denote the solution recovered by Spectral Experts. While not a perfect fit, it provides an good initialization for EM to further improve the solution (solid lines). (b) The dotted lines show different local optima found by EM.

Table 1. Parameter error $\|\theta^* - \hat{\theta}\|_F$ ($n = 500,000$) as the number of base variables b, number of features d and the number of components k increases. While Spectral by itself does not produce good parameter estimates, Spectral+EM improves over EM significantly.

Variables (b)	Features (d)	Components (k)	Spectral	EM	Spectral + EM
1	4	2	2.45 ± 3.68	0.28 ± 0.82	$\mathbf{0.17 \pm 0.57}$
2	5	2	1.38 ± 0.84	$\mathbf{0.00 \pm 0.00}$	$\mathbf{0.00 \pm 0.00}$
2	5	3	2.92 ± 1.71	0.43 ± 1.07	$\mathbf{0.31 \pm 1.02}$
2	6	2	2.33 ± 0.67	0.63 ± 1.29	$\mathbf{0.01 \pm 0.01}$



(a) Well-specified data

(b) Misspecified data

Figure 3. Learning curves: parameter error as a function of the number of samples n ($b = 1, d = 5, k = 3$).

Table 2. Parameter error $\|\theta^* - \hat{\theta}\|_F$ when the data is misspecified ($n = 500,000$). Spectral+EM degrades slightly, but still outperforms EM overall.

Variables (b)	Features (d)	Components (k)	Spectral	EM	Spectral + EM
1	4	2	1.70 ± 0.85	0.29 ± 0.85	$\mathbf{0.03 \pm 0.09}$
2	5	3	1.37 ± 0.85	0.44 ± 1.12	$\mathbf{0.00 \pm 0.00}$
2	6	5	9.89 ± 4.46	$\mathbf{2.53 \pm 1.77}$	2.69 ± 1.83
2	8	7	23.07 ± 7.10	9.62 ± 1.03	$\mathbf{8.16 \pm 2.31}$

feature set and 10 attempts for each instance. The experiments were run using $n = 500,000$ samples.

One of the main reasons for the high variance is the variation across random instances; some are easy for EM to find the global minima and others more difficult. In general, while Spectral Experts did not recover parameters by itself extremely well, it provided a good initialization for EM.

To study the stability of the solutions returned by Spectral Experts, consider the histogram in Figure 1, which shows the recovery errors of the algorithms over 170 attempts on a dataset with $b = 1, d = 4, k = 3$. Typically, Spectral Experts returned a stable solution. When these parameters were close enough to the true parameters, we found that EM almost always converged to the global optima. Randomly initialized EM only finds the true parameters a little over 10% of the time and shows considerably higher variance.

Effect of number of data points In Figure 3, we show how the recovery error varies as we get more data. Each data point shows the mean error over 10 attempts, with error bars. We note that the recovery performance of EM does not particularly improve; this suggests that EM continues to get stuck in a local optima. The spectral algorithm's error decays slowly, and as it gets closer to zero, EM initialized at the spectral parameters finds the true parameters more often as well. This behavior highlights the trade-off between statistical and computational error.

Misspecified data To evaluate how robust the algorithm was to model mis-specification, we removed large contiguous sections from $x \in [-0.5, -0.25] \cup [0.25, 0.5]$ and ran the algorithms again. Table 2 reports recovery errors in this scenario. The error in the estimates grows larger for higher d.

6. Conclusion

In this paper, we developed a computationally efficient and statistically consistent estimator for mixture of linear regressions. Our algorithm, Spectral Experts, regresses on higher-order powers of the data with a regularizer that encourages low rank structure, followed by tensor factorization to recover the actual parameters. Empirically, we found Spectral Experts to be an excellent initializer for EM.

Acknowledgements We would like to thank Lester Mackey for his fruitful suggestions and the anonymous reviewers for their helpful comments.

References

Anandkumar, A., Foster, D. P., Hsu, D., Kakade, S. M., and Liu, Y. Two SVDs suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2012a. MIT Press.

Anandkumar, A., Hsu, D., and Kakade, S. M. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory (COLT)*, 2012b.

Anandkumar, Anima, Ge, Rong, Hsu, Daniel, Kakade, Sham M., and Telgarsky, Matus. Tensor decompositions for learning latent variable models. *CoRR*, abs/1210.7559, 2012c.

Balle, B. and Mohri, M. Spectral learning of general weighted automata via constrained matrix completion. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2012. MIT Press.

Balle, B., Quattoni, A., and Carreras, X. A spectral learning algorithm for finite state transducers. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2011.

Candes, E. J., Strohmer, T., and Voroninski, V. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. Technical report, ArXiv, 2011.

Chaganty, A. and Liang, P. Spectral experts for estimating mixtures of linear regressions. *International Conference on Machine Learning*, 2013.

Cohen, S. B., Stratos, K., Collins, M., Foster, D. P., and Ungar, L. Spectral learning of latent-variable PCFGs. In *Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2012.

Grant, Michael and Boyd, Stephen. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, September 2012.

Hsu, D. and Kakade, S. M. Learning mixtures of spherical gaussians: Moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science (ITCS)*, 2013.

Hsu, D., Kakade, S. M., and Zhang, T. A spectral algorithm for learning hidden Markov models. In *Conference on Learning Theory (COLT)*, 2009.

Hsu, D., Kakade, S. M., and Liang, P. Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2012. MIT Press.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87, 1991.

Liang, P., Bouchard-Côté, A., Klein, D., and Taskar, B. An end-to-end discriminative approach to machine translation. In *International Conference on Computational Linguistics and Association for Computational Linguistics (COLING/ACL)*, Sydney, Australia, 2006. Association for Computational Linguistics.

Negahban, S. and Wainwright, M. J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *ArXiv e-prints*, December 2009.

Ohlsson, H., Yang, A., Dong, R., and Sastry, S. CPRL – an extension of compressive sensing to the phase retrieval problem. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2012. MIT Press.

Parikh, A., Song, L., Ishteva, M., Teodoru, G., and Xing, E. A spectral algorithm for latent junction trees. In *Uncertainty in Artificial Intelligence (UAI)*, 2012.

Petrov, S. and Klein, D. Discriminative log-linear grammars with latent variables. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2008. MIT Press.

Quattoni, A., Collins, M., and Darrell, T. Conditional random fields for object recognition. In *Advances in Neural Information Processing Systems (NIPS)*, Cambridge, MA, 2004. MIT Press.

Song, L., Boots, B., Siddiqi, S., Gordon, G., and Smola, A. Hilbert space embeddings of hidden Markov models. In *International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010. Omnipress.

Tomioka, R., Suzuki, T., Hayashi, K., and Kashima, H. Statistical performance of convex tensor decomposition. *Advances in Neural Information Processing Systems (NIPS)*, pp. 137, 2011.

Viele, Kert and Tong, Barbara. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12:315–330, 2002. ISSN 0960-3174. doi: 10.1023/A:1020779827503. URL http://dx.doi.org/10.1023/A%3A1020779827503.

Wang, Y. and Mori, G. Max-margin hidden conditional random fields for human action recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

A. Proofs: Regression

In this section, we will derive an upper bound on $\kappa(\mathfrak{X}_p)$ and a lower bound on $\frac{1}{n}\| \mathfrak{X}_p^*(\eta_p)\|_{\text{op}}$.

Lemma 5 (Lower bound on restricted strong convexity)**.** *Let* $\Sigma_p \stackrel{\text{def}}{=} \mathbb{E}[\text{cvec}(x^{\otimes p})^{\otimes 2}]$. *If*

$$n \geq \frac{16(p!)^2 R^{4p}}{\sigma_{\min}(\Sigma_p)^2} \left(1 + \sqrt{\frac{\log(1/\delta)}{2}}\right)^2,$$

then, with probability at least $1-\delta$,

$$\kappa(\mathfrak{X}_p) \geq \frac{\sigma_{\min}(\Sigma_p)}{2}.$$

Proof. Recall the definition of $\kappa(\mathfrak{X}_p)$,

$$\frac{1}{n}\| \mathfrak{X}_p(\Delta)\|_2^2 \geq \kappa(\mathfrak{X}_p)\|\Delta\|_F^2.$$

Expanding the definition of the observation operator:

$$\frac{1}{n}\| \mathfrak{X}_p(\Delta)\|_2^2 = \frac{1}{n}\sum_{(x,y)\in\mathcal{D}_p} \langle \Delta, x^{\otimes p}\rangle^2.$$

Unfolding the tensors, letting $\hat{\Sigma}_p \stackrel{\text{def}}{=} \frac{1}{n}\sum_{(x,y)\in\mathcal{D}_p} \text{cvec}(x^{\otimes p})^{\otimes 2}$, $\frac{1}{n}\| \mathfrak{X}_p(\Delta)\|_2^2 = \text{tr}(\text{cvec}(\Delta)^{\otimes 2}\hat{\Sigma}_p)$. We recall that each element of $\text{cvec}(\Delta)$ aggregates elements with permuted indices, so $\|\text{vec}(\Delta)\|_2 \leq \|\text{cvec}(\Delta)\|_2 \leq p!\|\text{vec}(\Delta)\|_2$. Then, we have

$$\begin{aligned}\frac{1}{n}\| \mathfrak{X}_p(\Delta)\|_2^2 &= \text{tr}(\text{cvec}(\Delta)^{\otimes 2}\hat{\Sigma}_p) \\ &\geq \sigma_{\min}(\hat{\Sigma}_p)\|\Delta\|_F^2.\end{aligned}$$

By Weyl's theorem,

$$\sigma_{\min}(\hat{\Sigma}_p) \geq \sigma_{\min}(\Sigma_p) - \|\hat{\Sigma}_p - \Sigma_p\|_{\text{op}}.$$

Since $\|\hat{\Sigma}_p - \Sigma_p\|_{\text{op}} \leq \|\hat{\Sigma}_p - \Sigma_p\|_F$, it suffices to show that the empirical covariance concentrates in Frobenius norm. Applying Lemma 8, with probability at least $1-\delta$,

$$\|\hat{\Sigma}_p - \Sigma_p\|_F \leq \frac{2\|\Sigma_p\|_F}{\sqrt{n}} \left(1 + \sqrt{\frac{\log(1/\delta)}{2}}\right).$$

Now we seek to control $\|\Sigma_p\|_F$. Since $\|x\|_2 \leq R$, we can use the bound

$$\|\Sigma_p\|_F \leq p!\|\text{vec}(x^{\otimes p})^{\otimes 2}\|_F \leq p! R^{2p}.$$

Finally, $\|\hat{\Sigma}_p - \Sigma_p\|_{\text{op}} \leq \sigma_{\min}(\Sigma_p)/2$ with probability at least $1-\delta$ if,

$$n \geq \frac{16(p!)^2 R^{4p}}{\sigma_{\min}(\Sigma_p)^2} \left(1 + \sqrt{\frac{\log(1/\delta)}{2}}\right)^2.$$

□

Lemma 6 (Upper bound on adjoint operator)**.** *With probability at least $1-\delta$, the following holds,*

$$\begin{aligned}
\frac{1}{n}\| \mathfrak{X}_1^*(\eta_1)\|_{\mathrm{op}} &\le \frac{2R(2LR+S)}{\sqrt{n}}\left(1+\sqrt{\frac{\log(3/\delta)}{2}}\right)\\
\frac{1}{n}\| \mathfrak{X}_2^*(\eta_2)\|_{\mathrm{op}} &\le \frac{(4L^2R^2+2SLR+4S^2)R^2}{\sqrt{n}}\left(1+\sqrt{\frac{\log(3/\delta)}{2}}\right)\\
\frac{1}{n}\| \mathfrak{X}_3^*(\eta_3)\|_{\mathrm{op}} &\le \frac{(8L^3R^3+3L^2R^2S+6LRS^2+2S^3)R^3}{\sqrt{n}}\left(1+\sqrt{\frac{\log(6/\delta)}{2}}\right)\\
&\quad + 3R^4S^2\left(\frac{4R(2LR+S)}{\sigma_{\min}(\Sigma_1)\sqrt{n}}\left(1+\sqrt{\frac{\log(6/\delta)}{2}}\right)\right).
\end{aligned}$$

It follows that, with probability at least $1-\delta$,

$$\frac{1}{n}\| \mathfrak{X}_p^*(\eta_p)\|_{\mathrm{op}} = O\left(L^pS^pR^{2p}\sigma_{\min}(\Sigma_1)^{-1}\sqrt{\frac{\log(1/\delta)}{n}}\right),$$

for each $p\in\{1,2,3\}$.

Proof. Let $\hat{\mathbb{E}}_p[f(x,\epsilon,h)]$ denote the empirical expectation over the examples in dataset $\mathcal{D}_p$ (recall the $\mathcal{D}_p$'s are independent to simplify the analysis). By definition,

$$\frac{1}{n}\| \mathfrak{X}_p^*(\eta_p)\|_{\mathrm{op}} = \left\|\hat{\mathbb{E}}_p[\eta_p(x)x^{\otimes p}]\right\|_{\mathrm{op}}$$

for $p\in\{1,2,3\}$. To proceed, we will bound each $\eta_p(x)$ and use Lemma 8 to bound $\|\hat{\mathbb{E}}_p[\eta_p(x)x^{\otimes p}]\|_F$. The Frobenius norm to bounds the operator norm, completing the proof.

Bounding $\eta_p(x)$. Using the assumptions that $\|\beta_h\|_2\le L$, $\|x\|_2\le R$ and $|\epsilon|\le S$, it is easy to bound each $\eta_p(x)$,

$$\begin{aligned}
\eta_1(x) &= \langle \beta_h - M_1, x\rangle + \epsilon\\
&\le \|\beta_h - M_1\|_2\|x\|_2 + |\epsilon|\\
&\le 2LR+S\\
\eta_2(x) &= \langle \beta_h^{\otimes 2} - M_2, x^{\otimes 2}\rangle + 2\epsilon\langle\beta_h, x\rangle + (\epsilon^2 - \mathbb{E}[\epsilon^2])\\
&\le \|\beta_h^{\otimes 2} - M_2\|_F\|x^{\otimes 2}\|_F + 2|\epsilon|\|\beta_h\|_2\|x\|_2 + |\epsilon^2-\mathbb{E}[\epsilon^2]|\\
&\le (2L)^2R^2 + 2SLR + (2S)^2\\
\eta_3(x) &= \langle \beta_h^{\otimes 3} - M_3, x^{\otimes 3}\rangle + 3\epsilon\langle\beta_h^{\otimes 2}, x^{\otimes 2}\rangle\\
&\quad + 3\left(\epsilon^2\langle\beta_h, x\rangle - \mathbb{E}[\epsilon^2]\langle\hat{M}_1, x\rangle\right) + (\epsilon^3 - \mathbb{E}[\epsilon^3])\\
&\le \|\beta_h^{\otimes 3} - M_3\|_F\|x^{\otimes 3}\|_F + 3|\epsilon|\|\beta_h^{\otimes 2}\|_F\|x^{\otimes 2}\|_F\\
&\quad + 3\left(|\epsilon^2|\ \|\beta_h\|_F\|x\|_F + \left|\mathbb{E}[\epsilon^2]\right|\ \|\hat{M}_1\|_2\|x\|_2\right) + |\epsilon^3| + \left|\mathbb{E}[\epsilon^3]\right|\\
&\le (2L)^3R^3 + 3SL^2R^2 + 3(S^2LR + S^2LR) + 2S^3.
\end{aligned}$$

We have used inequality $\|M_1-\beta_h\|_2\le 2L$ above.

Bounding $\left\|\hat{\mathbb{E}}[\eta_p(x)x^{\otimes p}]\right\|_F$. We may now apply the above bounds on $\eta_p(x)$ to bound $\|\eta_p(x)x^{\otimes p}\|_F$, using the fact that $\|cX\|_F \le c\|X\|_F$. By Lemma 8, each of the following holds with probability at least $1-\delta_1$,

$$\left\|\hat{\mathbb{E}}_1[\eta_1(x)x]\right\|_2 \le \frac{2R(2LR+S)}{\sqrt{n}}\left(1+\sqrt{\frac{\log(1/\delta_1)}{2}}\right)$$
$$\left\|\hat{\mathbb{E}}_2[\eta_2(x)x^{\otimes 2}]\right\|_F \le \frac{(4L^2R^2+2SLR+4S^2)R^2}{\sqrt{n}}\left(1+\sqrt{\frac{\log(1/\delta_2)}{2}}\right)$$
$$\left\|\hat{\mathbb{E}}_3[\eta_3(x)x^{\otimes 3}] - \mathbb{E}[\eta_3(x)x^{\otimes 3}\mid x]\right\|_F \le \frac{(8L^3R^3+3L^2R^2S+6LRS^2+2S^3)R^3}{\sqrt{n}}\left(1+\sqrt{\frac{\log(1/\delta_3)}{2}}\right).$$

Recall that $\eta_3(x)$ does not have zero mean, so we must bound the bias:

$$\begin{aligned}\|\mathbb{E}[\eta_3(x)x^{\otimes 3}\mid x]\|_F &= \|3\,\mathbb{E}[\epsilon^2]\langle M_1-\hat{M}_1, x\rangle x^{\otimes 3}\|_F \\ &\le 3\,\mathbb{E}[\epsilon^2]\|M_1-\hat{M}_1\|_2\|x\|_2\|x^{\otimes 3}\|_F.\end{aligned}$$

Note that in all of this, both $\hat{M}_1$ and M_1 are treated as constants. Further, by applying Lemma 1, we have a bound on $\|M_1-\hat{M}_1\|_2$. So, with probability at least $1-\delta_3$,

$$\|\mathbb{E}[\eta_3(x)x^{\otimes 3}\mid x]\|_F \le 3R^4S^2\left(\frac{4R(2LR+S)}{\sigma_{\min}(\Sigma_1)\sqrt{n}}\left(1+\sqrt{\frac{\log(1/\delta_3)}{2}}\right)\right).$$

Finally, taking $\delta_1=\delta/3, \delta_2=\delta/3, \delta_3=\delta/6$, and taking the union bound over the bounds for $p\in\{1,2,3\}$, we get our result. $\square$

B. Proofs: Tensor Decomposition

Once we have estimated the moments from the data through regression, we apply the robust tensor eigen-decomposition algorithm to recover the parameters, β_h and π. However, the algorithm is guaranteed to work only for symmetric matrices with (nearly) orthogonal eigenvectors, so as a first step, we will need to whiten the third-order moment tensor using the second moments. Once we get the eigenvalues and eigenvectors from this orthogonal tensor, we have to undo the transformation by applying an un-whitening step. In this section, we present error bounds for each step, and combine them to prove the following lemma,

Lemma 7 (Tensor Decomposition with Whitening)**.** *Let $M_3=\sum_{h=1}^k \pi_h\beta_h^{\otimes 3}$. Let $\|\hat{M}_2-M_2\|_{\text{op}}$ and $\|\hat{M}_3-M_3\|_{\text{op}}$ both be less than*

$$\frac{3\sigma_k(M_2)^{3/2}}{10k\pi_{\max}^{5/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)}+2\sqrt{2}\right)}\,\epsilon,$$

and,

$$\frac{\sigma_k(M_2)}{\|M_2\|_{\text{op}}^{1/2}\left(4\sqrt{3/2}+8k\pi_{\max}\sigma_k(M_2)^{-1/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)}+2\sqrt{2}\right)\right)}\,\epsilon,$$

for some ϵ such that

$$\epsilon \le \min \Bigg\{ \Bigg(4\sqrt{3/2} \|M_2\|_{\mathrm{op}}^{1/2} \sigma_k(M_2)^{-1} \varepsilon_{M_2} + 8\|M_2\|_{\mathrm{op}}^{1/2} k \pi_{\max} \sigma_k(M_2)^{-3/2} \left(24 \frac{\|M_3\|_{\mathrm{op}}}{\sigma_k(M_2)} + 2\sqrt{2} \right) \Bigg) \frac{\sigma_k(M_2)}{2}, \left(\frac{2\pi_{\max}^{3/2}}{3} 5k\pi_{\max}\sigma_k(M_2)^{-3/2} \left(24 \frac{\|M_3\|_{\mathrm{op}}}{\sigma_k(M_2)} + 2\sqrt{2} \right) \right) \frac{\sigma_k(M_2)}{2}, \frac{1}{2\sqrt{\pi_{\max}}} \Bigg\}.$$

Then, there exists a permutation of indices such that the parameter estimates found in step 2 of Algorithm 1 satisfy the following with probability at least $1 - \delta$,

$$\|\hat{\pi} - \pi\|_\infty \le \epsilon$$
$$\|\hat{\beta}_h - \beta_h\|_2 \le \epsilon.$$

for all $h \in [k]$.

Proof. We will use the general notation, $\varepsilon_X \stackrel{\mathrm{def}}{=} \|\hat{X} - X\|_{\mathrm{op}}$ to represent the error of the estimate, $\hat{X}$, of X in the operator norm.

Step 1: Whitening Let W and $\hat{W}$ be the whitening matrices for M_2 and $\hat{M}_2$ respectively. Also define $W^\dagger$ and $\hat{W}^\dagger$ to be their pseudo-inverses.

We will first show that the whitened tensors $T = M_3(W,W,W)$ and $\hat{T} = \hat{M}_3(\hat{W},\hat{W},\hat{W})$ are symmetric with orthogonal eigenvectors. Recall that $M_2 = \sum_h \pi_h \beta_h^{\otimes 2}$, and thus $W\beta_h = \frac{v_h}{\sqrt{\pi_h}}$, where v_h form an orthonormal basis. Applying the whitening transform to M_3, we get,

$$\begin{aligned} M_3 &= \sum_h \pi_h \beta_h^{\otimes 3} \\ M_3(W,W,W) &= \sum_h \pi_h (W\beta_h)^{\otimes 3} \\ &= \sum_h \frac{1}{\sqrt{\pi_h}} v_h^{\otimes 3}. \end{aligned}$$

Consequently, T has orthogonal eigenvectors, with eigenvalues $1/\sqrt{\pi_h}$.

Let us now study how far $\hat{T}$ differs from T, in terms of the errors of M_2 and M_3. To do so, we use the triangle inequality to break the difference into a number of simple terms,

$$\begin{aligned} \varepsilon_T &= \|M_3(W,W,W) - \hat{M}_3(\hat{W},\hat{W},\hat{W})\|_{\mathrm{op}} \\ &\le \|M_3(W,W,W) - M_3(W,W,\hat{W})\|_{\mathrm{op}} + \|M_3(W,W,\hat{W}) - M_3(W,\hat{W},\hat{W})\|_{\mathrm{op}} \\ &\quad + \|M_3(W,\hat{W},\hat{W}) - M_3(\hat{W},\hat{W},\hat{W})\|_{\mathrm{op}} + \|M_3(\hat{W},\hat{W},\hat{W}) - -\hat{M}_3(\hat{W},\hat{W},\hat{W})\|_{\mathrm{op}} \\ &\le \|M_3\|_{\mathrm{op}} \|W\|_{\mathrm{op}}^2 \varepsilon_W + \|M_3\|_{\mathrm{op}} \|\hat{W}\|_{\mathrm{op}} \|W\|_{\mathrm{op}} \varepsilon_W + \|M_3\|_{\mathrm{op}} \|\hat{W}\|_{\mathrm{op}}^2 \varepsilon_W + \varepsilon_{M_3} \|\hat{W}\|_{\mathrm{op}}^3 \\ &\le \|M_3\|_{\mathrm{op}} (\|W\|_{\mathrm{op}}^2 + \|\hat{W}\|_{\mathrm{op}} \|W\|_{\mathrm{op}} + \|\hat{W}\|_{\mathrm{op}}^2) \varepsilon_W + \varepsilon_{M_3} \|\hat{W}\|_{\mathrm{op}}^3 \end{aligned}$$

We can relate $\|\hat{W}\|$ and ε_W to ε_{M_2} using using Proposition 1. The conditions on ε_{M_2} imply that $\varepsilon_{M_2} < \sigma_k(M_2)/2$, giving us,

$$\begin{aligned} \|\hat{W}\|_{\mathrm{op}} &\le \sqrt{2}\sigma_k(M_2)^{-1/2} \\ \varepsilon_W &\le 4\sigma_k(M_2)^{-3/2} \varepsilon_{M_2}. \end{aligned}$$

Thus,

$$\begin{aligned}
\varepsilon_T &\leq 6\|M_3\|_{\text{op}}\|W\|_{\text{op}}^2(4\sigma_k(M_2)^{-3/2})\varepsilon_{M_2} + \varepsilon_{M_3}2\sqrt{2}\|W\|_{\text{op}}^3 \\
&\leq 24\|M_3\|_{\text{op}}\sigma_k(M_2)^{-5/2}\varepsilon_{M_2} + 2\sqrt{2}\sigma_k(M_2)^{-3/2}\varepsilon_{M_3} \\
&\leq \sigma_k(M_2)^{-3/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)} + 2\sqrt{2}\right)\max\{\varepsilon_{M_2}, \varepsilon_{M_3}\}.
\end{aligned}$$

Step 2: Decomposition We have constructed T to be a symmetric tensor with orthogonal eigenvectors. We can now apply the results of Anandkumar et al. (2012c, Theorem 5.1) to bound the error in the eigenvalues, λ_W, and eigenvectors, ω, returned by the robust tensor power method;

$$\begin{aligned}
\|\lambda_W - \hat{\lambda}_W\|_\infty &\leq \frac{5k\varepsilon_T}{(\lambda_W)_{\min}} \\
\|\omega_h - \hat{\omega}_h\|_2 &\leq \frac{8k\varepsilon_T}{(\lambda_W)_{\min}^2},
\end{aligned}$$

for all $h \in [k]$, where $(\lambda_W)_{min}$ is the smallest eigenvalue of T.

Step 3: Unwhitening Finally, we need to invert the whitening transformation to recover π and β_h from λ_W and ω_h. Let us complete the proof by studying how this inversion relates the error in π and β to the error in λ_W and ω.

First, we will bound the error in the βs,

$$\begin{aligned}
\|\hat{\beta}_h - \beta_h\|_2 &= \|\hat{W}^\dagger\hat{\omega} - W^\dagger\omega\|_2 \\
&\leq \varepsilon_{W^\dagger}\|\hat{\omega}_h\|_2 + \|W^\dagger\|_2\|\hat{\omega}_h - \omega_h\|_2.
\end{aligned}$$

Once more, we can apply the results of Proposition 1, with the assumptions on ε_{M_2}, to get,

$$\begin{aligned}
\|\hat{W}^\dagger\|_{\text{op}} &\leq \sqrt{3/2}\|M_2\|_{\text{op}}^{1/2} \\
\varepsilon_{W^\dagger} &\leq 4\sqrt{3/2}\|M_2\|_{\text{op}}^{1/2}\sigma_k(M_2)^{-1}\varepsilon_{M_2}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\|\hat{\beta}_h - \beta_h\|_2 &\leq 4\sqrt{3/2}\|M_2\|_{\text{op}}^{1/2}\sigma_k(M_2)^{-1}\varepsilon_{M_2} + 8\|M_2\|_{\text{op}}^{1/2}\frac{k\varepsilon_T}{(\lambda_W)_{min}^2} \\
&\leq 4\sqrt{3/2}\|M_2\|_{\text{op}}^{1/2}\sigma_k(M_2)^{-1}\varepsilon_{M_2} \\
&\quad + 8\|M_2\|_{\text{op}}^{1/2}k\pi_{\max}\sigma_k(M_2)^{-3/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)} + 2\sqrt{2}\right)\max\{\varepsilon_{M_2}, \varepsilon_{M_3}\}.
\end{aligned}$$

Next, let us bound the error in π,

$$\begin{aligned}
|\hat{\pi}_h - \pi_h| &= \left|\frac{1}{(\lambda_W)_h^2} - \frac{1}{(\hat{\lambda}_W)_h^2}\right| \\
&= \left|\frac{\left((\lambda_W)_h + (\hat{\lambda}_W)_h\right)\left((\lambda_W)_h - (\hat{\lambda}_W)_h\right)}{(\lambda_W)_h^2(\hat{\lambda}_W)_h^2}\right| \\
&\leq \frac{(2(\lambda_W)_h - \|\lambda_W - \hat{\lambda}_W\|_\infty)}{(\lambda_W)_h^2\left((\lambda_W)_h + \|\lambda_W - \hat{\lambda}_W\|_\infty\right)^2}\|\lambda_W - \hat{\lambda}_W\|_\infty.
\end{aligned}$$

Recall that $(\lambda_W)_h = \pi_h^{-1/2}$, so the assumptions that ϵ imply that $\|\lambda_W - \hat{\lambda}_W\|_\infty \leq (\lambda_W)_{\min}/2$. This allows us to simplify the above expression as follows,

$$\begin{aligned}
|\hat{\pi}_h - \pi_h| &\leq \frac{(3/2)(\lambda_W)_h}{(3/2)^2(\lambda_W)_h^4}\|\lambda_W - \hat{\lambda}_W\|_\infty \\
&\leq \frac{2}{3(\lambda_W)_h^3}\frac{5k\varepsilon_T}{(\lambda_W)_{\min}^2} \\
&\leq \frac{2\pi_{\max}^{3/2}}{3} 5k\pi_{\max}\sigma_k(M_2)^{-3/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)} + 2\sqrt{2}\right)\max\{\varepsilon_{M_2}, \varepsilon_{M_3}\}.
\end{aligned}$$

We complete the proof by requiring that the bounds ε_{M_2} and ε_{M_3} imply that $\|\hat{\pi} - \pi\|_\infty \leq \epsilon$ and $\|\hat{\beta}_h - \beta_h\|_2 \leq \epsilon$, i.e.

$$\max\{\varepsilon_{M_2}, \varepsilon_{M_3}\} \leq \frac{3\sigma_k(M_2)^{3/2}}{10k\pi_{\max}^{5/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)} + 2\sqrt{2}\right)}\ \epsilon,$$

as well as,

$$\max\{\varepsilon_{M_2}, \varepsilon_{M_3}\} \leq \frac{\sigma_k(M_2)}{\|M_2\|_{\text{op}}^{1/2}\left(4\sqrt{3/2} + 8k\pi_{\max}\sigma_k(M_2)^{-1/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)} + 2\sqrt{2}\right)\right)}\ \epsilon.$$

□

C. Basic Lemmas

Lemma 8 (Concentration of vector norms)**.** *Let $X, X_1, \cdots, X_n \in \mathbb{R}^d$ be i.i.d. samples from some distribution with bounded support ($\|X\|_2 \leq M$ with probability 1). Then with probability at least $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]\right\|_2 \leq \frac{2M}{\sqrt{n}}\left(1 + \sqrt{\frac{\log(1/\delta)}{2}}\right).$$

Proof. Define $Z_i = X_i - \mathbb{E}[X]$.

The quantity we want to bound can be expressed as follows:

$$f(Z_1, Z_2, \cdots, Z_n) = \left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\|_2.$$

Let us check that f satisfies the bounded differences inequality:

$$\begin{aligned}
|f(Z_1, \cdots, Z_i, \cdots, Z_n) - f(Z_1, \cdots, Z_i', \cdots, Z_n)| &\leq \frac{1}{n}\|Z_i - Z_i'\|_2 \\
&= \frac{1}{n}\|X_i - X_i'\|_2 \\
&\leq \frac{2M}{n},
\end{aligned}$$

by the bounded assumption of X_i and the triangle inequality.

By McDiarmid's inequality, with probability at least $1-\delta$, we have:

$$\Pr[f - \mathbb{E}[f] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (2M/n)^2}\right).$$

Re-arranging:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_2 \leq \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\|_2 \right] + M\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Now it remains to bound $\mathbb{E}[f]$. By Jensen's inequality, $\mathbb{E}[f] \leq \sqrt{\mathbb{E}[f^2]}$, so it suffices to bound $\mathbb{E}[f^2]$:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n^2} \left\| \sum_{i=1}^{n} Z_i \right\|^2 \right] &= \mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^{n} \|Z_i\|_2^2 \right] + \mathbb{E}\left[\frac{1}{n^2} \sum_{i \neq j} \langle Z_i, Z_j \rangle \right] \\
&\leq \frac{4M^2}{n} + 0,
\end{aligned}$$

where the cross terms are zero by independence of the Z_i's.

Putting everything together, we obtain the desired bound:

$$\left\| \frac{1}{n} \sum_{i=1}^{n} Z_i \right\| \leq \frac{2M}{\sqrt{n}} + M\sqrt{\frac{2\log(1/\delta)}{n}}.$$

□

Remark: The above result can be directly applied to the Frobenius norm of a matrix M because $\|M\|_F = \|\operatorname{vec}(M)\|_2$.

Proposition 1 (Perturbation Bounds on Whitening Matrices)**.** *Let A be a rank-k $d \times d$ matrix, $\hat{W}$ be a $d \times k$ matrix that whitens $\hat{A}$, i.e. $\hat{W}^T \hat{A} \hat{W} = I$. Suppose $\hat{W}^T A \hat{W} = UDU^T$, then define $W = \hat{W} U D^{-\frac{1}{2}} U^T$. Note that W is also a $d \times k$ matrix that whitens A. Let $\alpha_A = \frac{\varepsilon_A}{\sigma_k(A)}$.*

Then,

$$\begin{aligned}
\|\hat{W}\|_{\mathrm{op}} &\leq \frac{\|W\|_{\mathrm{op}}}{\sqrt{1-\alpha_A}} \\
\|\hat{W}^\dagger\|_{\mathrm{op}} &\leq \|W^\dagger\|_{\mathrm{op}} \sqrt{1+\alpha_A} \\
\varepsilon_W &\leq 2\|W\|_{\mathrm{op}} \frac{\alpha_A}{1-\alpha_A} \\
\varepsilon_{W^\dagger} &\leq 2\|W^\dagger\|_{\mathrm{op}} \sqrt{1+\alpha_A} \frac{\alpha_A}{1-\alpha_A}.
\end{aligned}$$

Proof. First, note that for a matrix W that whitens $A = V\Sigma V^T$, $W = V\Sigma^{-\frac{1}{2}} V^T$ and $W^\dagger = V\Sigma^{-\frac{1}{2}} V^T$. This allows us to bound the operator norms of $\hat{W}$ and $\hat{W}^\dagger$ in terms of W and $W^\dagger$,

$$\begin{aligned}
\|\hat{W}\|_{\mathrm{op}} &= \frac{1}{\sqrt{\sigma_k(\hat{A})}} \\
&\leq \frac{1}{\sqrt{\sigma_k(A) - \varepsilon_A}} \\
&\leq \frac{\|W\|_{\mathrm{op}}}{\sqrt{1-\alpha_A}} \\
\|\hat{W}^\dagger\|_{\mathrm{op}} &= \sqrt{\sigma_1(\hat{A})} \\
&\leq \sqrt{\sigma_{\max}(A) + \varepsilon_A} \\
&\leq \sqrt{1+\alpha_A}\|W^\dagger\|_{\mathrm{op}}.
\end{aligned}$$

To find ε_W, we will exploit the rotational invariance of the operator norm.

$$\begin{aligned}
\varepsilon_W &= \|\hat{W} - W\|_{\text{op}} \\
&= \|WUD^{\frac{1}{2}}U^T - W\|_{\text{op}} \\
&\le \|W\|_{\text{op}}\|I - UD^{\frac{1}{2}}U^T\|_{\text{op}} \\
&\le \|W\|_{\text{op}}\|I - D\|_{\text{op}} \\
&= \|W\|_{\text{op}}\|I - UDU^T\|_{\text{op}} \\
&\le \|W\|_{\text{op}}\|\hat{W}^T\hat{A}_k\hat{W} - \hat{W}^T A\hat{W}\|_{\text{op}} \\
&\le \|W\|_{\text{op}}(\|\hat{W}^T(\hat{A}_k - \hat{A})\hat{W}\|_{\text{op}} + \|\hat{W}^T(\hat{A} - A)\hat{W}\|_{\text{op}}) \\
&\le \|W\|_{\text{op}}\|\hat{W}\|_{\text{op}}^2(\sigma_{k+1}(\hat{A}) + \varepsilon_A) \\
&\le 2\|W\|_{\text{op}}\|\hat{W}\|_{\text{op}}^2\varepsilon_A \\
&\le 2\|W\|_{\text{op}}\frac{\alpha_A}{1 - \alpha_A}.
\end{aligned}$$

Similarly, we can bound the error on the un-whitening transform, $W^\dagger$,

$$\begin{aligned}
\varepsilon_{W^\dagger} &= \|\hat{W}^\dagger - W^\dagger\|_{\text{op}} \\
&= \|\hat{W}^\dagger UD^{\frac{1}{2}}U^T - W^\dagger\|_{\text{op}} \\
&\le \|\hat{W}^\dagger\|_{\text{op}}\|I - UD^{\frac{1}{2}}U^T\|_{\text{op}} \\
&\le 2\|\hat{W}^\dagger\|_{\text{op}}\|\hat{W}\|_{\text{op}}^2\varepsilon_A \\
&\le 2\|W^\dagger\|_{\text{op}}\sqrt{1 + \alpha_A}\frac{\alpha_A}{1 - \alpha_A}.
\end{aligned}$$

□