Journal of Machine Learning Research 1–11

Supplementary Material for Spectral Experts for Estimating Mixtures of Linear Regressions

Arun Tejasvi Chaganty CHAGANTY@CS.STANFORD.EDU
Percy Liang PLIANG@CS.STANFORD.EDU

1. Review of Notation

Let $[n] = \{1, \dots, n\}$ denote the first n positive integers.

We use $x^{\otimes p}$ to represent the p-th order tensor formed by taking the outer 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: $\langle X, Y \rangle = \sum_{i_1,\dots i_p} X_{i_1,\dots i_p} Y_{i_1,\dots,i_p}$. 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), let $\|X\|_F$ denote the Frobenius norm (square root of sum of squares of singular values), let $\|X\|_{\max}$ denote the max norm (elementwise maximum), let $\|X\|_{\text{op}}$ denote the operator norm (largest singular value), let $\sigma_{\min}(X)$ be the smallest singular value of X. For a 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 symmetric tensor $X \in (\mathbb{R}^d)^{\otimes p}$, let $\text{cvec}(X) \in \mathbb{R}^{C_{d,p}}, C_{d,p} = \binom{d+p+1}{p}$ be the collapsed vectorization of distinct elements in X, for example, for $X \in \mathbb{R}^{d \times d}$, $\text{cvec}(X) = (X_{ii} : i \in [d]; X_{ij} + X_{ji} : i, j \in [d], i < j)$. 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 with that count profile.

2. Regression

Let us review the regression problem set up in (Chaganty and Liang, 2013, Section 3). We assume we are given data $(x_i, y_i) \in \mathcal{D}_p$ generated by the following process,

$$y_i = \langle M_p, x_i^{\otimes p} \rangle + \eta_p(x_i), \tag{1}$$

where $M_p = \sum_{h=1}^k \pi_h \beta_h^{\otimes p}$, the p-th order moments of β_h and $\eta_p(x)$ is zero mean noise. In particular, for $p \in \{1,2,3\}$, we showed that $\eta_p(x)$ were defined to be,

$$\eta_1(x) = \langle \beta_h - M_1, x \rangle + \epsilon \tag{2}$$
$$\eta_2(x) = \langle \beta_h^{\otimes 2} - M_2, x^{\otimes 2} \rangle + 2\epsilon \langle \beta_h, x \rangle + (\epsilon^2 - \mathbb{E}[\epsilon^2]) \tag{3}$$
$$\eta_3(x) = \langle \beta_h^{\otimes 3} - M_3, x^{\otimes 3} \rangle + 3\epsilon \langle \beta_h^{\otimes 2}, x^{\otimes 2} \rangle + 3(\epsilon^2 \langle \beta_h, x \rangle - \mathbb{E}[\epsilon^2]\langle M_1, x \rangle) + (\epsilon^3 - \mathbb{E}[\epsilon^3]). \tag{4}$$

We assume that $\|x_i\| \le R$, $\|\beta_h\| \le L$ and $|\epsilon| \le S$.

We then defined the observation operator $\mathfrak{X}_p(M_p) : \mathbb{R}^{d^{\otimes p}} \to \mathbb{R}^n$,

$$\mathfrak{X}_p(M_p; \mathcal{D}_p)_i \stackrel{\text{def}}{=} \langle M_p, x_i^{\otimes p} \rangle, \qquad (x_i, y_i) \in \mathcal{D}_p, \tag{5}$$

which let us succinctly represent the low-rank regression problem as follows,

$$\min_{M_p \in \mathbb{R}^{d^{\otimes p}}} \frac{1}{2n} \|y - \mathfrak{X}_p(M_p; \mathcal{D}_p)\|_2^2 + \lambda_p \|M_p\|_*. \tag{6}$$

Let us also recall the adjoint of the observation operator, $\mathfrak{X}_p^* : \mathbb{R}^n \to \mathbb{R}^{d^p}$,

$$\mathfrak{X}_p^*(\eta_p; \mathcal{D}_p) = \sum_{x \in \mathcal{D}_p} \eta_p(x) x^{\otimes p}, \tag{7}$$

where we have used η_p to represent the vector $[\eta_p(x)]_{x \in \mathcal{D}_p}$.

Tomioka et al. (2011) showed that error in the estimated $\hat{M}_p$ can be bounded as follows;

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

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

Then the error of $\hat{M}_p$ is bounded as follows: $\|\hat{M}_p - M_p^*\|_F \le \frac{\lambda_n \sqrt{k}}{\kappa(\mathfrak{X}_p)}$.

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

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

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

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

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

$$\begin{aligned} \frac{1}{n}\| \mathfrak{X}_p(\Delta)\|_2^2 &= \mathrm{tr}(\mathrm{cvec}(\Delta)^{\otimes 2}\hat{\Sigma}_p) && (8)\\ &\geq \sigma_{\min}(\hat{\Sigma}_p)\|\Delta\|_F^2. && (9)\end{aligned}$$

By Weyl's theorem,

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

Since $\|\hat{\Sigma}_p - \Sigma_p\|_{\mathrm{op}} \leq \|\hat{\Sigma}_p - \Sigma_p\|_F$, it suffices to show that the empirical covariance concentrates in Frobenius norm. Applying Lemma 5, 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

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

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

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

∎

Lemma 3 (Upper bound on adjoint operator) *With probability at least $1-\delta$, the following holds,*

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

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

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

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 $\eta_p(x)$, defined in (2), (3) and (4) and use Lemma 5 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 $\eta_p(x)$,

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

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 5, each of the following holds with probability at least $1-\delta_1$,

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

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

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

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 $1 - \delta_3$,

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

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. ∎

3. 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 4 (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

$$\epsilon \leq \min \Bigg\{ \Bigg(4\sqrt{3/2}\|M_2\|_{\text{op}}^{1/2}\sigma_k(M_2)^{-1}\varepsilon_{M_2} \tag{10}$$

$$+ 8\|M_2\|_{\text{op}}^{1/2} k\pi_{\max}\sigma_k(M_2)^{-3/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)} + 2\sqrt{2}\right)\Bigg)\frac{\sigma_k(M_2)}{2}, \tag{11}$$

$$\left(\frac{2\pi_{\max}^{3/2}}{3}5k\pi_{\max}\sigma_k(M_2)^{-3/2}\left(24\frac{\|M_3\|_{\text{op}}}{\sigma_k(M_2)} + 2\sqrt{2}\right)\right)\frac{\sigma_k(M_2)}{2}, \tag{12}$$

$$\frac{1}{2\sqrt{\pi_{\max}}} \tag{13}$$

$$\Bigg\}. \tag{14}$$

Then, there exists a permutation of indices such that the parameter estimates found in step 2 of Chaganty and Liang (2013, 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 $h \in [k]$.

Proof We will use the general notation, $\varepsilon_X \stackrel{\text{def}}{=} \|\hat{X} - X\|_{\text{op}}$ 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,

$$M_3 = \sum_h \pi_h \beta_h^{\otimes 3} \tag{15}$$

$$M_3(W, W, W) = \sum_h \pi_h (W\beta_h)^{\otimes 3} \tag{16}$$

$$= \sum_h \frac{1}{\sqrt{\pi_h}} v_h^{\otimes 3}. \tag{17}$$

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,

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

We can relate $\|\hat{W}\|$ and ε_W to ε_{M_2} using using Proposition 6. The conditions on ε_{M_2} imply that $\varepsilon_{M_2} < \sigma_k(M_2)/2$, giving us,

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

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. (2012, Theorem 5.1) to bound the error in the eigenvalues, λ_W, and eigenvectors, ω, returned by the robust tensor power method;

$$\|\lambda_W - \hat{\lambda}_W\|_\infty \leq \frac{5k\varepsilon_T}{(\lambda_W)_{\min}} \tag{18}$$

$$\|\omega_h - \hat{\omega}_h\|_2 \leq \frac{8k\varepsilon_T}{(\lambda_W)_{\min}^2}, \tag{19}$$

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,

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

Once more, we can apply the results of Proposition 6, 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,

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

Next, let us bound the error in π,

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

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

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

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

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

as well as,

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

■

4. Basic Lemmas

Lemma 5 (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 ($\|X\|_2 \le M$ with probability 1). Then with probability at least $1-\delta$,*

$$\left\| \frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}[X] \right\|_2 \le \frac{2M}{\sqrt{n}} \left(1 + \sqrt{\frac{\log(1/\delta)}{2}} \right). \tag{20}$$

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

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

$$f(Z_1, Z_2, \cdots, Z_n) = \left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\|_2. \tag{21}$$

Let us check that f satisfies the bounded differences inequality:

$$\begin{aligned} |f(Z_1, \cdots, Z_i, \cdots, Z_n) - f(Z_1, \cdots, Z_i', \cdots, Z_n)| &\le \frac{1}{n} \|Z_i - Z_i'\|_2 && (22) \\ &= \frac{1}{n} \|X_i - X_i'\|_2 && (23) \\ &\le \frac{2M}{n}, && (24) \end{aligned}$$

by the bounded assumption of X_i and the triangle inequality.

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

$$\mathbb{P}[f - \mathbb{E}[f] \ge \epsilon] \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (2M/n)^2} \right). \tag{25}$$

Re-arranging:

$$\left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\|_2 \le \mathbb{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\|_2 \right] + M \sqrt{\frac{2\log(1/\delta)}{n}}. \tag{26}$$

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

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n^2} \left\| \sum_{i=1}^n Z_i \right\|^2 \right] &= \mathbb{E}\left[\frac{1}{n^2} \sum_{i=1}^n \|Z_i\|_2^2 \right] + \mathbb{E}\left[\frac{1}{n^2} \sum_{i \ne j} \langle Z_i, Z_j \rangle \right] && (27) \\ &\le \frac{4M^2}{n} + 0, && (28) \end{aligned}$$

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

Putting everything together, we obtain the desired bound:

$$\left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\| \le \frac{2M}{\sqrt{n}} + M \sqrt{\frac{2\log(1/\delta)}{n}}. \tag{29}$$

$\blacksquare$

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

Proposition 6 (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,

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

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 $W^\dagger$,

$$\begin{aligned}
\|\hat{W}\|_{\text{op}} &= \frac{1}{\sqrt{\sigma_k(\hat{A})}} \\
&\le \frac{1}{\sqrt{\sigma_k(A) - \varepsilon_A}} && \text{(By Weyl's Theorem)} \\
&\le \frac{\|W\|_{\text{op}}}{\sqrt{1-\alpha_A}} \\
\|\hat{W}^\dagger\|_{\text{op}} &= \sqrt{\sigma_1(\hat{A})} \\
&\le \sqrt{\sigma_{\max}(A) + \varepsilon_A} && \text{(By Weyl's Theorem)} \\
&\le \sqrt{1+\alpha_A} \|W^\dagger\|_{\text{op}}.
\end{aligned}$$

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

$$\begin{aligned}
\varepsilon_W &= \|\hat{W} - W\|_{\mathrm{op}} \\
&= \|WUD^{\frac{1}{2}}U^T - W\|_{\mathrm{op}} && (W = UD^{-\frac{1}{2}}U^T) \\
&\le \|W\|_{\mathrm{op}}\|I - UD^{\frac{1}{2}}U^T\|_{\mathrm{op}} && \text{(Sub-multiplicativity)} \\
&\le \|W\|_{\mathrm{op}}\|I - D\|_{\mathrm{op}} \\
&= \|W\|_{\mathrm{op}}\|I - UDU^T\|_{\mathrm{op}} && \text{(Rotational invariance)} \\
&\le \|W\|_{\mathrm{op}}\|\hat{W}^T\hat{A}_k\hat{W} - \hat{W}^T A\hat{W}\|_{\mathrm{op}} && \text{(By definition)} \\
&\le \|W\|_{\mathrm{op}}(\|\hat{W}^T(\hat{A}_k - \hat{A})\hat{W}\|_{\mathrm{op}} + \|\hat{W}^T(\hat{A} - A)\hat{W}\|_{\mathrm{op}}) \\
&\le \|W\|_{\mathrm{op}}\|\hat{W}\|_{\mathrm{op}}^2(\sigma_{k+1}(\hat{A}) + \varepsilon_A) \\
&\le 2\|W\|_{\mathrm{op}}\|\hat{W}\|_{\mathrm{op}}^2\varepsilon_A && \text{(Since } \sigma_{k+1}(A) = 0) \\
&\le 2\|W\|_{\mathrm{op}}\frac{\alpha_A}{1 - \alpha_A} \quad . && \text{(Using bound on } \|\hat{W}\|_{\mathrm{op}})
\end{aligned}$$

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

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

■

References

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

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

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