BACKGROUND OF sLDA AND TENSOR DECOMPOSITION

Supervised LDA (Blei and McAuliffe, 2007)

\[
\alpha \rightarrow h \rightarrow z \rightarrow x \rightarrow \mu \\
\eta, \sigma \rightarrow y \\
(N, \beta)
\]

(a) \( y_d = \eta_d^T h_d + \varepsilon_d \)

- Data: Document collection \( x_{dtr} \), response variables \( y_d \in \mathbb{R} \).
- \( O = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^{V \times k} \) : the topic dictionary.
- \( h_d \in \Delta^{k-1} \) : topic mixing vectors for each document.
- \( z_{dn} \in \{1, \ldots, k\} \) : topic assignments for each word.
- \( \eta \in \mathbb{R}^k \) : the linear regression model.
- \( \alpha \in \mathbb{R}^k \) : prior parameter for topic mixing vectors.

Orthogonal tensor decomposition

For a \( p \)-th order tensor \( T \), find orthonormal basis \( \{v_i\}_{i=1}^p \) and scalars \( \{\lambda_i\}_{i=1}^p \) such that \( T = \sum_{i=1}^p \lambda_i v_i \otimes \cdots \otimes v_i \).

Tools: robust tensor power method (Anandkumar et al., 2012), ALS method.

Question: Can we use tensor decomposition based methods to obtain consistent estimates of sLDA parameters?

THE LEARNING ALGORITHM AND SAMPLING COMPLEXITY ANALYSIS

1. **Input**: Document collection \( x_{dtr} \), response variables \( y_d \) and \( \alpha_0 = \|\alpha\|_1 \).
2. Compute empirical moments and obtain \( \hat{M}_2, \hat{M}_3 \) and \( \hat{M}_y \).
3. Find \( W \in \mathbb{R}^{v \times k} \) such that \( \hat{M}_3(W, W) = \mathbf{I}_k \).
4. Find robust eigenvalues and eigenvectors \((\hat{\lambda}_i, \hat{v}_i)\) of \( \hat{M}_3(W, W, W) \) using the robust tensor power method (Anandkumar et al., 2012).
5. Recover parameters: \( \hat{\alpha}_i \leftarrow \frac{4(\alpha_0 + 1)}{(\alpha_0 + 2)^2} \hat{\lambda}_i \hat{v}_i^T \), \( \hat{\mu}_i \leftarrow \frac{4\alpha_0 + 2}{2} \hat{\lambda}_i \hat{v}_i^T \hat{M}_y(W, W) \hat{v}_i \).
6. **Output**: \( \hat{\eta}, \hat{\alpha} \) and \( \{\hat{\mu}_i\}_{i=1}^k \).

THEOREM (SAMPLE COMPLEXITY ANALYSIS)

Suppose we have an sLDA model \( M = \{\mu_i\}_{i=1}^k, \alpha, \eta\) and \( n \) documents i.i.d. sampled from \( M \). Let \( O = (\mu_1, \ldots, \mu_k) \). Then the estimates \( \hat{M} = \{\hat{\mu}_i\}_{i=1}^k, \hat{\alpha}, \hat{\eta}\) is \( \epsilon \)-close to \( M \) with high probability given that \( n \geq \Omega(\text{poly}(\epsilon, k, \sigma_k(O)^{-1}, \alpha_{\text{max}}, \alpha_{\text{min}}^{-1} \|\eta\|)) \).

EXPERIMENTS

**Synthetic datasets**

![Parameter estimation error (1-norm) on synthetic datasets.](image)

**Amazon movie review dataset (McAuley and Leskovec, 2013)**

![pR² scores (top) and negative per-word log-likelihood (bottom) on Amazon movie review dataset.](image)

**Table**: Running time for Gibbs sampling and spectral method.

<table>
<thead>
<tr>
<th>( k )</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n \times 10^4 )</td>
<td>1</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Gibbs-sLDA</td>
<td>0.6</td>
<td>3.0</td>
<td>6.0</td>
</tr>
<tr>
<td>Spec-sLDA</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
</tr>
</tbody>
</table>

yiningwa@cs.cmu.edu  dcszj@mail.tsinghua.edu.cn