Learning latent variable models using tensor decompositions

Daniel Hsu

Computer Science Department & Data Science Institute
Columbia University

Machine Learning Summer School
June 29-30, 2018
Learning algorithms
for latent variable models
based on decompositions of moment tensors.
Learning algorithms (parameter estimation) for latent variable models based on decompositions of moment tensors.

“Method-of-moments” (Pearson, 1894)
Example #1: summarizing a corpus of documents

Observation: **documents express one or more thematic topics.**

*Politics Ensnare Mohamed Salah and Switzerland at the World Cup*

By Rory Smith, James Montague and Tariq Panja

June 24, 2018

MOSCOW — The World Cup was thrust into the combustible mix of politics and soccer — dangerous ground that world soccer takes great pains to avoid — as a growing number of disciplinary proceedings and a star player’s threatened retirement brought several sensitive international flash points to the tournament’s doorstep this weekend.
Example #1: summarizing a corpus of documents

Observation: documents express one or more thematic topics.

Politics Ensnare Mohamed Salah and Switzerland at the World Cup
By Rory Smith, James Montague and Tariq Panja
June 24, 2018

MOSCOW — The World Cup was thrust into the combustible mix of politics and soccer — dangerous ground that world soccer takes great pains to avoid — as a growing number of disciplinary proceedings and a star player’s threatened retirement brought several sensitive international flash points to the tournament’s doorstep this weekend.

- What topics are expressed in a corpus of documents?
- How prevalent is each topic in the corpus?
Topic model (e.g., latent Dirichlet allocation)

$K$ topics (distributions over vocab words).

Document $\equiv$ mixture of topics.

Word tokens in doc. $\sim$ iid mixture distribution.
Topic model (e.g., latent Dirichlet allocation)

$K$ topics (distributions over vocab words).

Document $\equiv$ mixture of topics.

Word tokens in doc. \( \text{iid} \) mixture distribution.

E.g.,

\[ \text{iid} \sim 0.7 \times P_{\text{sports}} + 0.3 \times P_{\text{politics}}. \]
Topic model (e.g., latent Dirichlet allocation)

$K$ topics (distributions over vocab words).

Document $\equiv$ mixture of topics.

Word tokens in doc. $\sim$ mixture distribution.

E.g., $\sim 0.7 \times P_{\text{sports}} + 0.3 \times P_{\text{politics}}$.

Given corpus of documents (and “hyper-parameters”, e.g., $K$), produce estimates of model parameters, e.g.:

- Distribution $P_t$ over vocab words, for each $t \in [K]$.
- Weight $w_t$ of topic $t$ in document corpus, for each $t \in [K]$. 
Suppose each word token $x$ in document is annotated with source topic $t_x \in \{1, 2, \ldots, K\}$.

<table>
<thead>
<tr>
<th>Politics</th>
<th>Ensnare</th>
<th>Mohamed_Salah</th>
<th>and</th>
<th>Switzerland</th>
<th>at</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>
Suppose each word token \( x \) in document is annotated with source topic \( t_x \in \{1, 2, \ldots, K\} \).

<table>
<thead>
<tr>
<th>Politics</th>
<th>Ensnare</th>
<th>Mohamed_Salah</th>
<th>and</th>
<th>Switzerland</th>
<th>at</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

Then estimating the \( \{ (P_t, w_t) \}_{t=1}^K \) can be done “directly”.
Suppose each word token \( x \) in document is *annotated* with source topic \( t_x \in \{1, 2, \ldots, K\} \).

Then estimating the \( \{(P_t, w_t)\}_{t=1}^K \) can be done “directly”.

Unfortunately, we often don’t have such annotations (i.e., data are *unlabeled* / topics are *hidden*).

“Direct” approach to estimation unavailable.
Example #2: subpopulations in data

Data studied by Pearson (1894):
ratio of forehead-width to body-length for 1000 crabs.
Example #2: subpopulations in data

Data studied by Pearson (1894):
ratio of forehead-width to body-length for 1000 crabs.

Sample may be comprised of different sub-species of crabs.
Gaussian mixture model

\[ H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K); \]

\[ X \mid H = t \sim \text{Normal}(\mu_t, \Sigma_t), \quad t \in [K]. \]
Gaussian mixture model

\[ H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K); \]
\[ X \mid H = t \sim \text{Normal}(\mu_t, \Sigma_t), \quad t \in [K]. \]

Estimate \textit{mean vector}, \textit{covariance matrix}, and \textit{mixing weight} of each subpopulation from \textit{unlabeled data}. 
Maximum likelihood estimation

- No “direct” estimators when some variables are hidden.
Maximum likelihood estimation

- No “direct” estimators when some variables are hidden.

- **Maximum likelihood estimator** (MLE):

\[ \theta_{\text{MLE}} := \arg \max_{\theta \in \Theta} \log \Pr_{\theta} (\text{data}). \]
Maximum likelihood estimation

- No “direct” estimators when some variables are hidden.

- **Maximum likelihood estimator (MLE):**

\[
\theta_{\text{MLE}} := \arg \max_{\theta \in \Theta} \log \Pr_{\theta} (\text{data}) .
\]

- **Note:** log-likelihood is not necessarily concave function of \(\theta\).
Maximum likelihood estimation

- No “direct” estimators when some variables are hidden.

- **Maximum likelihood estimator (MLE):**

\[
\theta_{\text{MLE}} := \arg \max_{\theta \in \Theta} \log \Pr_{\theta} (\text{data}).
\]

- **Note:** log-likelihood is not necessarily concave function of \( \theta \).

- For latent variable models, often use local optimization, most notably via **Expectation-Maximization (EM)** (Dempster, Laird, & Rubin, 1977).
MLE for Gaussian mixture models

Given data \( \{ \mathbf{x}_i \}_{i=1}^{n} \), find \( \{ (\mu_t, \Sigma_t, \pi_t) \}_{t=1}^{K} \) to maximize

\[
\sum_{i=1}^{n} \log \left( \sum_{t=1}^{K} \pi_t \cdot \frac{1}{\det(\Sigma_t)^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \mu_t)^\top \Sigma_t^{-1} (\mathbf{x}_i - \mu_t) \right\} \right).
\]
MLE for Gaussian mixture models

Given data \( \{x_i\}_{i=1}^n \), find \( \{(\mu_t, \Sigma_t, \pi_t)\}_{t=1}^K \) to maximize

\[
\sum_{i=1}^n \log \left( \sum_{t=1}^K \frac{\pi_t}{\det(\Sigma_t)^{1/2}} \exp \left\{ -\frac{1}{2} (x_i - \mu_t)^\top \Sigma_t^{-1} (x_i - \mu_t) \right\} \right).
\]

- Sensible with restrictions on \( \Sigma_t \) (e.g., \( \Sigma_t \succeq \sigma^2 I \)).
MLE for Gaussian mixture models

Given data $\{x_i\}_{i=1}^n$, find $\{(\mu_t, \Sigma_t, \pi_t)\}_{t=1}^K$ to maximize

$$
\sum_{i=1}^n \log \left( \sum_{t=1}^K \pi_t \cdot \frac{1}{\det(\Sigma_t)^{1/2}} \exp \left\{ -\frac{1}{2} (x_i - \mu_t)^\top \Sigma_t^{-1} (x_i - \mu_t) \right\} \right).
$$

- Sensible with restrictions on $\Sigma_t$ (e.g., $\Sigma_t \succeq \sigma^2 I$).
- But NP-hard to maximize (Tosh and Dasgupta, 2018):
  
  Can’t expect efficient algorithms to work for all data sets.
Parameter learning objective

Suppose iid sample of size \( n \) is generated by distribution from model with (unknown) parameters \( \theta \in \Theta \subseteq \mathbb{R}^p \). (\( p = \# \) params)
Parameter learning objective

Suppose iid sample of size $n$ is generated by distribution from model with (unknown) parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. ($p = \# \text{ params}$)

**Task:** Produce estimate $\hat{\theta}$ of $\theta$ such that

$$\mathbb{E} \| \hat{\theta} - \theta \| \to 0 \quad \text{as} \quad n \to \infty$$

(i.e., $\hat{\theta}$ is *consistent*).
Parameter learning objective

Suppose iid sample of size $n$ is generated by distribution from model with (unknown) parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. ($p = \# \text{ params}$)

**Task**: Produce estimate $\hat{\theta}$ of $\theta$ such that

$$\mathbb{E} \|\hat{\theta} - \theta\| \to 0 \quad \text{as} \quad n \to \infty$$

(i.e., $\hat{\theta}$ is consistent).

- E.g., for spherical Gaussian mixtures:
  - For $K = 2$ (and $\pi_t = 1/2$, $\Sigma_t = I$): EM is consistent (Xu, H., & Maleki, 2016).
Parameter learning objective

Suppose iid sample of size $n$ is generated by distribution from model with (unknown) parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. ($p = \#\text{ params}$)

**Task**: Produce estimate $\hat{\theta}$ of $\theta$ such that

$$
\mathbb{E} \| \hat{\theta} - \theta \| \to 0 \quad \text{as} \quad n \to \infty
$$

(i.e., $\hat{\theta}$ is consistent).

- E.g., for spherical Gaussian mixtures:
  - For $K = 2$ (and $\pi_t = 1/2$, $\Sigma_t = I$): EM is consistent (Xu, H., & Maleki, 2016).
  - Larger $K$: easily trapped in local maxima, far from global max (Jin, Zhang, Balakrishnan, Wainwright, & Jordan, 2016).
Parameter learning objective

Suppose iid sample of size $n$ is generated by distribution from model with (unknown) parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. ($p = \# \text{ params}$)

**Task**: Produce estimate $\hat{\theta}$ of $\theta$ such that

$$\mathbb{E} \| \hat{\theta} - \theta \| \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty$$

(i.e., $\hat{\theta}$ is *consistent*).

- E.g., for spherical Gaussian mixtures:
  - For $K = 2$ (and $\pi_t = 1/2$, $\Sigma_t = I$): EM is consistent (Xu, H., & Maleki, 2016).
  - Larger $K$: easily trapped in local maxima, far from global max (Jin, Zhang, Balakrishnan, Wainwright, & Jordan, 2016).

Practitioners often use EM with many (random) restarts . . . but may take a long time to get near the global max.
Parameter learning objective

Suppose iid sample of size $n$ is generated by distribution from model with (unknown) parameters $\theta \in \Theta \subseteq \mathbb{R}^p$. ($p = \# \text{ params}$)

**Task**: Produce estimate $\hat{\theta}$ of $\theta$ such that

$$\Pr\left(\|\hat{\theta} - \theta\| \leq \epsilon\right) \geq 1 - \delta$$

with $\text{poly}(p, 1/\epsilon, 1/\delta, \ldots)$ sample size and running time.

- E.g., for spherical Gaussian mixtures:
  - For $K = 2$ (and $\pi_t = 1/2$, $\Sigma_t = I$): EM is consistent (Xu, H., & Maleki, 2016).
  - Larger $K$: easily trapped in local maxima, far from global max (Jin, Zhang, Balakrishnan, Wainwright, & Jordan, 2016).

Practitioners often use EM with many (random) restarts ... but may take a long time to get near the global max.
Hard to learn model parameters, even when data is generated by a model distribution.
Barriers

Hard to learn model parameters, even when data is generated by a model distribution.

Cryptographic hardness
(e.g., Mossel & Roch, 2006)

Information-theoretic hardness
(e.g., Moitra & Valiant, 2010)

May require $2^{\Omega(K)}$ running time or $2^{\Omega(K)}$ sample size.
Ways around the barriers

- **Separation conditions.**

  E.g., assume mixture component distributions are far apart. 
  (Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; ... )
Ways around the barriers

▶ **Separation conditions.**

E.g., assume mixture component distributions are far apart.
(Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; ...)

▶ **Structural assumptions.**

E.g., sparsity, anchor words.
(Spielman, Wang, & Wright, 2012; Arora, Ge, & Moitra, 2012; ...)

Ways around the barriers

- **Separation conditions.**
  
  E.g., assume mixture component distributions are far apart. 
  (Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; ...)

- **Structural assumptions.**
  
  E.g., sparsity, anchor words. 
  (Spielman, Wang, & Wright, 2012; Arora, Ge, & Moitra, 2012; ...)

- **Non-degeneracy conditions.**
  
  E.g., assume $\mu_1, \mu_2, \ldots, \mu_K$ are in general position.
Ways around the barriers

- **Separation conditions.**
  E.g., assume mixture component distributions are far apart. (Dasgupta, 1999; Arora & Kannan, 2001; Vempala & Wang, 2002; ...)

- **Structural assumptions.**
  E.g., sparsity, anchor words. (Spielman, Wang, & Wright, 2012; Arora, Ge, & Moitra, 2012; ...)

- **Non-degeneracy conditions.**
  E.g., assume $\mu_1, \mu_2, \ldots, \mu_K$ are in general position.

**This lecture:** learning algorithms for non-degenerate instances via *method-of-moments.*
Method-of-moments at a glance

1. Determine function of model parameters $\theta$ estimatable from observable data:

$$\mathbb{E}_\theta[f(X)] \quad ("moments").$$

2. Form estimates of moments using data (e.g., iid sample):

$$\hat{\mathbb{E}}[f(X)] \quad ("empirical \ moments").$$

3. Approximately solve equations for parameters $\theta$:

$$\mathbb{E}_\theta[f(X)] = \hat{\mathbb{E}}[f(X)].$$

4. ("Fine-tune" estimated parameters with local optimization.)
Method-of-moments at a glance

1. Determine function of model parameters $\theta$ estimatable from observable data:

$$\mathbb{E}_{\theta}[f(X)] \quad (\text{“moments”}).$$

Which moments?

2. Form estimates of moments using data (e.g., iid sample):

$$\widehat{\mathbb{E}}[f(X)] \quad (\text{“empirical moments”}).$$

3. Approximately solve equations for parameters $\theta$:

$$\mathbb{E}_{\theta}[f(X)] = \widehat{\mathbb{E}}[f(X)].$$

How?

4. (“Fine-tune” estimated parameters with local optimization.)
Method-of-moments at a glance

1. Determine function of model parameters $\theta$ estimatable from observable data:

$$\mathbb{E}_\theta[f(X)] \quad (\text{"moments"}).$$

Which moments? Often low-order moments suffice.

2. Form estimates of moments using data (e.g., iid sample):

$$\hat{\mathbb{E}}[f(X)] \quad (\text{"empirical moments"}).$$

3. Approximately solve equations for parameters $\theta$:

$$\mathbb{E}_\theta[f(X)] = \hat{\mathbb{E}}[f(X)].$$


4. ("Fine-tune" estimated parameters with local optimization.)
A simple example of the method-of-moments

Let $X \sim \text{Normal}(\mu, \sigma^2)$. How to estimate $\sigma^2$ from iid sample?
A simple example of the method-of-moments

Let $X \sim \text{Normal}(\mu, \sigma^2)$. How to estimate $\sigma^2$ from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$. 

- Formula for $\sigma^2$ in terms of moments:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2.$$ 

- Form estimates of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ from iid sample $\{x_i\}_{i=1}^n$:

$$\mathbb{E}[X] := \frac{1}{n} \sum_{i=1}^n x_i,$$
$$\mathbb{E}[X^2] := \frac{1}{n} \sum_{i=1}^n x_i^2.$$ 

- Then estimate $\sigma^2$ with $\hat{\sigma}^2 := \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

We'll follow this same basic recipe for much richer models!
A simple example of the method-of-moments

Let $X \sim \text{Normal}(\mu, \sigma^2)$. How to estimate $\sigma^2$ from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
- Formula for $\sigma^2$ in terms of moments:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \left(\sigma^2 + \mu^2\right) - \mu^2 = \sigma^2.$$
A simple example of the method-of-moments

Let $X \sim \text{Normal}(\mu, \sigma^2)$. How to estimate $\sigma^2$ from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
- Formula for $\sigma^2$ in terms of moments:
  \[ \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \left( \sigma^2 + \mu^2 \right) - \mu^2 = \sigma^2. \]

- Form estimates of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ from iid sample $\{x_i\}_{i=1}^n$:
  e.g.,
  \[ \hat{\mathbb{E}}[X] := \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\mathbb{E}}[X^2] := \frac{1}{n} \sum_{i=1}^n x_i^2. \]
A simple example of the method-of-moments

Let \( X \sim \text{Normal}(\mu, \sigma^2) \). How to estimate \( \sigma^2 \) from iid sample?

- Consider first- and second-moments: \( \mathbb{E}[X] \) and \( \mathbb{E}[X^2] \).
- Formula for \( \sigma^2 \) in terms of moments:

  \[
  \mathbb{E}[X^2] - \mathbb{E}[X]^2 = (\sigma^2 + \mu^2) - \mu^2 = \sigma^2.
  \]

- Form estimates of \( \mathbb{E}[X] \) and \( \mathbb{E}[X^2] \) from iid sample \( \{x_i\}_{i=1}^n \):
  e.g.,

  \[
  \hat{\mathbb{E}}[X] := \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\mathbb{E}}[X^2] := \frac{1}{n} \sum_{i=1}^n x_i^2.
  \]

- Then estimate \( \sigma^2 \) with

  \[
  \hat{\sigma}^2 := \hat{\mathbb{E}}[X^2] - \hat{\mathbb{E}}[X]^2.
  \]
A simple example of the method-of-moments

Let $X \sim \text{Normal}(\mu, \sigma^2)$. How to estimate $\sigma^2$ from iid sample?

- Consider first- and second-moments: $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$.
- Formula for $\sigma^2$ in terms of moments:

$$\mathbb{E}[X^2] - \mathbb{E}[X]^2 = \left( \sigma^2 + \mu^2 \right) - \mu^2 = \sigma^2.$$

- Form estimates of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ from iid sample $\{x_i\}_{i=1}^n$:

  e.g.,

  $$\hat{\mathbb{E}}[X] := \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\mathbb{E}}[X^2] := \frac{1}{n} \sum_{i=1}^n x_i^2.$$

- Then estimate $\sigma^2$ with

$$\hat{\sigma}^2 := \hat{\mathbb{E}}[X^2] - \hat{\mathbb{E}}[X]^2.$$

We’ll follow this same basic recipe for much richer models!
1. Topic model for single-topic documents.
   - Identifiability.
   - Parameter recovery via orthogonal tensor decomposition.

2. Moment decompositions for other models.
   - Mixtures of Gaussians and linear regressions.
   - Multi-view models (e.g., HMMs).
   - Other models (e.g., single-index models).

3. Error analysis.
Other models amenable to moment tensor decomposition

- Models for independent components analysis (Comon, 1994; Frieze, Jerrum, & Kannan, 1996; Arora, Ge, Moitra & Sachdeva, 2012; Anandkumar, Foster, H., Kakade, & Liu, 2012, 2015; Belkin, Rademacher, & Voss, 2013; etc.)
- Mixed-membership stochastic blockmodels (Anandkumar, Ge, H., & Kakade, 2013, 2014)
- Simple probabilistic grammars (H., Kakade, & Liang, 2012)
- Noisy-or networks (Halpern & Sontag, 2013; Jernite, Halpern & Sontag, 2013; Arora, Ge, Ma, & Risteski, 2016)
- Indian buffet process (Tung & Smola, 2014)
- Mixed multinomial logit model (Oh & Shah, 2014)
- Dawid-Skene model (Zhang, Chen, Zhou, & Jordan, 2014)
- Multi-task bandits (Azar, Lazaric, & Brunskill, 2013)
- Partially obs. MDPs (Azizzadenesheli, Lazaric, & Anandkumar, 2016)
- …
1. Topic model for single-topic documents
Topic model

General topic model (e.g., Latent Dirichlet Allocation)

\[ K \text{ topics (dists. over words)} \{ P_t \}_{t=1}^K. \]

Document \( \equiv \) mixture of topics \( \text{(hidden)}. \)

Word tokens in doc. \( \sim \) iid mixture distribution.
Topic model

**Topic model for single-topic documents**

- \( K \) topics (dists. over words) \( \{P_t\}_{t=1}^K \).
- Pick topic \( t \) with prob. \( w_t \) (hidden).
- Word tokens in doc. \( \sim \) iid \( P_t \).
Topic model

Topic model for single-topic documents

$K$ topics (dists. over words) $\{P_t\}_{t=1}^K$.
Pick topic $t$ with prob. $w_t$ (hidden).
Word tokens in doc. $\sim P_t$.

Given iid sample of documents of length $L$,
produce estimates of model parameters $\{(P_t, w_t)\}_{t=1}^K$. 
Topic model

Topic model for single-topic documents

$K$ topics (dists. over words) $\{P_t\}_{t=1}^K$.

Pick topic $t$ with prob. $w_t$ (hidden).

Word tokens in doc. $\sim P_t$.

Given iid sample of documents of length $L$,
produce estimates of model parameters $\{(P_t, w_t)\}_{t=1}^K$.

How long must the documents be?
Topic model

Topic model for single-topic documents

$K$ topics (dists. over words) $\{P_t\}_{t=1}^K$.

Pick topic $t$ with prob. $w_t$ (hidden).

Word tokens in doc. $\sim P_t$.

Given iid sample of documents of length $L$, produce estimates of model parameters $\{(P_t, w_t)\}_{t=1}^K$.

How long must the documents be?

(Answering this question leads to efficient algorithms for estimating parameters!)
Identifiability

**Generative process:**
Pick $t \sim \text{Categorical}(w_1, w_2, \ldots, w_K)$.
Given $t$, pick $L$ words from $P_t$. 
Identifiability

Generative process:
Pick $t \sim \text{Categorical}(w_1, w_2, \ldots, w_K)$.
Given $t$, pick $L$ words from $P_t$.

- $L = 1$: random document (single word) $\sim \sum_{t=1}^{K} w_t P_t$.

Are parameters $\{(P_t, w_t)\}_{t=1}^{K}$ identifiable from single-word documents?
Identifiability

**Generative process:**
Pick \( t \sim \text{Categorical}(w_1, w_2, \ldots, w_K) \).
Given \( t \), pick \( L \) words from \( P_t \).

- \( L = 1 \): random document (single word) \( \sim \sum_{t=1}^{K} w_t P_t \).

Are parameters \( \{(P_t, w_t)\}_{t=1}^{K} \) identifiable from single-word documents?
No.
Identifiability

**Generative process:**
Pick \( t \sim \text{Categorical}(w_1, w_2, \ldots, w_K) \).
Given \( t \), pick \( L \) words from \( P_t \).

- \( L = 1 \): random document (single word) \( \sim \sum_{t=1}^{K} w_t P_t \).

Are parameters \( \{(P_t, w_t)\}_{t=1}^{K} \) identifiable from single-word documents?
No.
Identifiability

Generative process:
Pick $t \sim \text{Categorical}(w_1, w_2, \ldots, w_K)$.
Given $t$, pick $L$ words from $P_t$.

- $L = 1$: random document (single word) \( \sim \sum_{t=1}^{K} w_t P_t \).

Are parameters \( \{(P_t, w_t)\}_{t=1}^{K} \) identifiable from single-word documents?

No.
Identifiability: \( L = 2 \)

**Generative process:**
Pick \( t \sim \text{Categorical}(w_1, w_2, \ldots, w_K) \).
Given \( t \), pick \( L \) words from \( P_t \).

- \( L = 2 \):
Identifiability: \( L = 2 \)

**Generative process:**
Pick \( t \sim \text{Categorical}(w_1, w_2, \ldots, w_K) \).
Given \( t \), pick \( L \) words from \( P_t \).

- \( L = 2 \):
  Regard \( P_t \) as probability vector (\( i \)th entry of \( P_t \) is \( \Pr[\text{word } i] \)).
  Joint distribution of word pairs (for topic \( t \)) is given by matrix:

\[
P_t P_t^\top = \begin{bmatrix}
  \Pr[\text{words } i, j]
end{bmatrix}
\]

Random document \( \sim \sum_{t=1}^{K} w_t P_t P_t^\top \).
Identifiability: \( L = 2 \)

**Generative process:**
Pick \( t \sim \text{Categorical}(w_1, w_2, \ldots, w_K) \).
Given \( t \), pick \( L \) words from \( P_t \).

- \( L = 2 \):
  Regard \( P_t \) as probability vector (\( i \)th entry of \( P_t \) is \( \Pr[\text{word } i] \)).
  Joint distribution of word pairs (for topic \( t \)) is given by matrix:

\[
P_t P_t^\top = \begin{pmatrix}
  \Pr[\text{words } i, j]
\end{pmatrix}
\]

Random document \( \sim \sum_{t=1}^{K} w_t P_t P_t^\top \).

Are parameters \( \{(P_t, w_t)\}_{t=1}^{K} \) identifiable from word pairs?
Simple observation

Suppose distribution of word pairs (as a matrix) can be written as

\[ M = AA^\top. \]
Simple observation

Suppose distribution of word pairs (as a matrix) can be written as

\[ M = AA^\top. \]

Then it can also be written as

\[ M = (AR)(AR)^\top \]

for any orthogonal matrix \( R \) (because \( R^\top R = I \)).
Identifiability: \( L = 2 \) counterexample

Parameters \( \{(P_1, w_1), (P_2, w_2)\} \) and \( \{(	ilde{P}_1, \tilde{w}_1), (\tilde{P}_2, \tilde{w}_2)\} \)

\[
(P_1, w_1) = \left( \begin{bmatrix} 0.40 \\ 0.60 \end{bmatrix}, 0.5 \right), \quad (P_2, w_2) = \left( \begin{bmatrix} 0.60 \\ 0.40 \end{bmatrix}, 0.5 \right);
\]

\[
(\tilde{P}_1, \tilde{w}_1) = \left( \begin{bmatrix} 0.55 \\ 0.45 \end{bmatrix}, 0.8 \right), \quad (\tilde{P}_2, \tilde{w}_2) = \left( \begin{bmatrix} 0.30 \\ 0.70 \end{bmatrix}, 0.2 \right).
\]
Identifiability: \( L = 2 \) counterexample

Parameters \( \{(P_1, w_1), (P_2, w_2)\} \) and \( \{(	ilde{P}_1, \tilde{w}_1), (\tilde{P}_2, \tilde{w}_2)\} \)

\[
(P_1, w_1) = \begin{pmatrix} 0.40 \\ 0.60 \end{pmatrix}, 0.5, \quad (P_2, w_2) = \begin{pmatrix} 0.60 \\ 0.40 \end{pmatrix}, 0.5;
\]

\[
(\tilde{P}_1, \tilde{w}_1) = \begin{pmatrix} 0.55 \\ 0.45 \end{pmatrix}, 0.8, \quad (\tilde{P}_2, \tilde{w}_2) = \begin{pmatrix} 0.30 \\ 0.70 \end{pmatrix}, 0.2
\]

satisfy

\[
w_1 P_1 P_1^\top + w_2 P_2 P_2^\top = \tilde{w}_1 \tilde{P}_1 \tilde{P}_1^\top + \tilde{w}_2 \tilde{P}_2 \tilde{P}_2^\top = \begin{pmatrix} 0.26 & 0.24 \\ 0.24 & 0.26 \end{pmatrix}.
\]
Identifiability: $L = 2$ counterexample

Parameters $\{(P_1, w_1), (P_2, w_2)\}$ and $\{(	ilde{P}_1, \tilde{w}_1), (\tilde{P}_2, \tilde{w}_2)\}$

\[
(P_1, w_1) = \begin{pmatrix} 0.40 \\ 0.60 \end{pmatrix}, \quad (P_2, w_2) = \begin{pmatrix} 0.60 \\ 0.40 \end{pmatrix};
\]

\[
(\tilde{P}_1, \tilde{w}_1) = \begin{pmatrix} 0.55 \\ 0.45 \end{pmatrix}, \quad (\tilde{P}_2, \tilde{w}_2) = \begin{pmatrix} 0.30 \\ 0.70 \end{pmatrix};
\]

satisfy

\[
w_1 P_1 P_1^\top + w_2 P_2 P_2^\top = \tilde{w}_1 \tilde{P}_1 \tilde{P}_1^\top + \tilde{w}_2 \tilde{P}_2 \tilde{P}_2^\top = \begin{pmatrix} 0.26 & 0.24 \\ 0.24 & 0.26 \end{pmatrix}.
\]

Cannot identify parameters from length-two documents.
Identifiability: $L = 3$

Documents of length $L = 3$
Joint distribution of word triple (for topic $t$) is given by tensor:

$$P_t \otimes P_t \otimes P_t = \text{Pr}[\text{words } i, j, k]$$

Random document $\sim \sum_{t=1}^{K} w_t P_t \otimes P_t \otimes P_t.$
Claim: If $\{P_t\}_{t=1}^K$ are linearly independent & all $w_t > 0$, then parameters $\{(P_t, w_t)\}_{t=1}^K$ are identifiable from word triples.
Claim: If \( \{P_t\}_{t=1}^K \) are linearly independent & all \( w_t > 0 \), then parameters \( \{(P_t, w_t)\}_{t=1}^K \) are identifiable from word triples.

Claim implied by uniqueness of certain tensor decompositions.
Claim: If \( \{P_t\}_{t=1}^K \) are linearly independent & all \( w_t > 0 \), then parameters \( \{ (P_t, w_t) \}_{t=1}^K \) are identifiable from word triples.

- Claim implied by uniqueness of certain tensor decompositions.
- Proof is constructive: i.e., comes with an algorithm!
Claim: If \( \{P_t\}_{t=1}^K \) are linearly independent & all \( w_t > 0 \), then parameters \( \{ (P_t, w_t) \}_{t=1}^K \) are identifiable from word triples.

▶ Claim implied by uniqueness of certain tensor decompositions.
▶ Proof is constructive: i.e., comes with an algorithm!

Next: Brief overview of tensors.
Tensors of order two

Matrices (tensors of order two): \( M \in \mathbb{R}^{d \times d} \).

- Regard as \emph{bi-linear function} \( M : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \):

\[
M(ax + x', y) = aM(x, y) + M(x', y) ;
\]
\[
M(x, ay + y') = aM(x, y) + M(x, y') .
\]
Tensors of order two

Matrices (tensors of order two): $M \in \mathbb{R}^{d \times d}$.

- Regard as bi-linear function $M : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:
  
  \[ M(\alpha x + x', y) = \alpha M(x, y) + M(x', y) ; \]
  
  \[ M(x, \alpha y + y') = \alpha M(x, y) + M(x, y') . \]

- Can describe $M$ by $d^2$ values $M(e_i, e_j) =: M_{i,j}$.

  ($e_i$ is $i$th coordinate basis vector.)
Matrices (tensors of order two): $M \in \mathbb{R}^{d \times d}$.

- Regard as bi-linear function $M : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$:

  $M(ax + x', y) = aM(x, y) + M(x', y)$;
  $M(x, ay + y') = aM(x, y) + M(x, y')$.

- Can describe $M$ by $d^2$ values $M(e_i, e_j) =: M_{i,j}$.
  ($e_i$ is $i$th coordinate basis vector.)

- Formula using matrix representation:

  $M(x, y) = x^\top My = \sum_{i,j} M_{i,j} \cdot x_i y_j$.
Tensors of order two

Matrices (tensors of order two): $M \in \mathbb{R}^{d \times d}$.

- Regard as bi-linear function $M : \mathbb{R}^{d} \times \mathbb{R}^{d} \rightarrow \mathbb{R}$:

\[
M(ax + x', y) = aM(x, y) + M(x', y); \\
M(x, ay + y') = aM(x, y) + M(x, y').
\]

- Can describe $M$ by $d^2$ values $M(e_i, e_j) =: M_{i,j}$.
  
  ($e_i$ is $i$th coordinate basis vector.)

- Formula using matrix representation:

\[
M(x, y) = x^\top My = \sum_{i,j} M_{i,j} \cdot x_i y_j.
\]

Tensors are multi-linear generalization.
Tensors of order $p$

$p$-linear functions: $T: \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$. 
Tensors of order $p$

$p$-linear functions: $T: \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \rightarrow \mathbb{R}$.

- Can describe $T$ by $d^p$ values $T(e_{i_1}, e_{i_2}, \ldots, e_{i_p}) =: T_{i_1, i_2, \ldots, i_p}$. 

$26$
Tensors of order $p$

$p$-linear functions: $T : \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Can describe $T$ by $d^p$ values $T(e_{i_1}, e_{i_2}, \ldots, e_{i_p}) =: T_{i_1, i_2, \ldots, i_p}$.

- Identify $T$ with multi-index array $T \in \mathbb{R}^{d \times d \times \cdots \times d}$. 

26
Tensors of order \( p \)

\( p \)-linear functions: \( T: \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R} \).

- Can describe \( T \) by \( d^p \) values \( T(e_{i_1}, e_{i_2}, \ldots, e_{i_p}) =: T_{i_1,i_2,\ldots,i_p} \).
- Identify \( T \) with multi-index array \( T \in \mathbb{R}^{d \times d \times \cdots \times d} \).

Formula for function value:

\[
T(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \sum_{i_1,i_2,\ldots,i_p} T_{i_1,i_2,\ldots,i_p} \cdot x^{(1)}_{i_1} x^{(2)}_{i_2} \cdots x^{(p)}_{i_p}.
\]
Tensors of order $p$

$p$-linear functions: $T: \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \rightarrow \mathbb{R}$.

- Can describe $T$ by $d^p$ values $T(e_{i_1}, e_{i_2}, \ldots, e_{i_p}) =: T_{i_1,i_2,\ldots,i_p}$.

- Identify $T$ with multi-index array $T \in \mathbb{R}^{d \times d \times \cdots \times d}$.

Formula for function value:

$$T(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \sum_{i_1,i_2,\ldots,i_p} T_{i_1,i_2,\ldots,i_p} \cdot x^{(1)}_{i_1} x^{(2)}_{i_2} \cdots x^{(p)}_{i_p}.$$  

- Rank-1 tensor: $T = v^{(1)} \otimes v^{(2)} \otimes \cdots \otimes v^{(p)}$,

$$T(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \langle v^{(1)}, x^{(1)} \rangle \langle v^{(2)}, x^{(2)} \rangle \cdots \langle v^{(p)}, x^{(p)} \rangle.$$
Tensors of order $p$

$p$-linear functions: $T : \mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$.

- Can describe $T$ by $d^p$ values $T(e_{i_1}, e_{i_2}, \ldots, e_{i_p}) =: T_{i_1, i_2, \ldots, i_p}$.
- Identify $T$ with multi-index array $T \in \mathbb{R}^{d \times d \times \cdots \times d}$.

Formula for function value:

$$T(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \sum_{i_1, i_2, \ldots, i_p} T_{i_1, i_2, \ldots, i_p} \cdot x^{(1)}_{i_1} x^{(2)}_{i_2} \cdots x^{(p)}_{i_p}.$$  

- **Rank-1 tensor**: $T = v^{(1)} \otimes v^{(2)} \otimes \cdots \otimes v^{(p)}$,

$$T(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \langle v^{(1)}, x^{(1)} \rangle \langle v^{(2)}, x^{(2)} \rangle \cdots \langle v^{(p)}, x^{(p)} \rangle.$$  

**Symmetric rank-1 tensor**: $T = v^{\otimes p} = v \otimes v \otimes \cdots \otimes v$,

$$T(x^{(1)}, x^{(2)}, \ldots, x^{(p)}) = \langle v, x^{(1)} \rangle \langle v, x^{(2)} \rangle \cdots \langle v, x^{(p)} \rangle.$$
Most Tensor Problems Are NP-Hard

CHRISTOPHER J. HILLAR, Mathematical Sciences Research Institute
LEK-HENG LIM, University of Chicago

We prove that multilinear (tensor) analogues of many efficiently computable problems in numerical linear algebra are NP-hard. Our list includes: determining the feasibility of a system of bilinear equations, deciding whether a 3-tensor possesses a given eigenvalue, singular value, or spectral norm; approximating an eigenvalue, eigenvector, singular vector, or the spectral norm; and determining the rank or best rank-1 approximation of a 3-tensor. Furthermore, we show that restricting these problems to symmetric tensors does not alleviate their NP-hardness. We also explain how deciding nonnegative definiteness of a symmetric 4-tensor is NP-hard and how computing the combinatorial hyperdeterminant is NP-, #P-, and VNP-hard.
Example: rank

- Rank of $T$: smallest $r$ s.t. $T$ is sum of $r$ rank-1 tensors.
Example: rank

- Rank of $T$: smallest $r$ s.t. $T$ is sum of $r$ rank-1 tensors.
  (Computing this is NP-hard.)
Example: rank

- Rank of $T$: smallest $r$ s.t. $T$ is sum of $r$ rank-1 tensors. (Computing this is NP-hard.)

- “Border rank” of $T$: smallest $r$ s.t. there exists sequence $(T_k)_{k \in \mathbb{N}}$ of rank-$r$ tensors with $\lim_{k \to \infty} T_k = T$. 

Define $T = x \otimes x \otimes y + x \otimes y \otimes x + y \otimes x \otimes x$, which has rank 3.

Define $T_1/\epsilon := 1/\epsilon (x + \epsilon y) \otimes (x + \epsilon y) \otimes (x + \epsilon y) - 1/\epsilon x \otimes x \otimes x$, which have rank 2. For $\epsilon = 1/k$, have $\lim_{k \to \infty} T_k = T$. 

28
Example: rank

- Rank of $T$: smallest $r$ s.t. $T$ is sum of $r$ rank-1 tensors. (Computing this is NP-hard.)

- “Border rank” of $T$: smallest $r$ s.t. there exists sequence $(T_k)_{k \in \mathbb{N}}$ of rank-$r$ tensors with $\lim_{k \to \infty} T_k = T$.

- Rank is not same as border rank!

Define

$$T := x \otimes x \otimes y + x \otimes y \otimes x + y \otimes x \otimes x,$$

which has rank 3.
Example: rank

- Rank of $T$: smallest $r$ s.t. $T$ is sum of $r$ rank-1 tensors. (Computing this is NP-hard.)

- “Border rank” of $T$: smallest $r$ s.t. there exists sequence $(T_k)_{k \in \mathbb{N}}$ of rank-$r$ tensors with $\lim_{k \to \infty} T_k = T$.

- Rank is not same as border rank!

Define

$$T := x \otimes x \otimes y + x \otimes y \otimes x + y \otimes x \otimes x,$$

which has rank 3.

Define

$$T_{1/\epsilon} := \frac{1}{\epsilon} (x + \epsilon y) \otimes (x + \epsilon y) \otimes (x + \epsilon y) - \frac{1}{\epsilon} x \otimes x \otimes x,$$

which have rank 2.
Example: rank

- Rank of $T$: smallest $r$ s.t. $T$ is sum of $r$ rank-1 tensors. (Computing this is NP-hard.)

- “Border rank” of $T$: smallest $r$ s.t. there exists sequence $(T_k)_{k \in \mathbb{N}}$ of rank-$r$ tensors with $\lim_{k \to \infty} T_k = T$.

- Rank is not same as border rank!

Define

$$T := x \otimes x \otimes y + x \otimes y \otimes x + y \otimes x \otimes x,$$

which has rank 3.

Define

$$T_{1/\epsilon} := \frac{1}{\epsilon}(x + \epsilon y) \otimes (x + \epsilon y) \otimes (x + \epsilon y) - \frac{1}{\epsilon} x \otimes x \otimes x,$$

which have rank 2.

For $\epsilon = 1/k$, have $\lim_{k \to \infty} T_k = T$. 
Aside: eigenvalue decomposition

**Recall:** every symmetric matrix $M \in \mathbb{R}^{d \times d}$ of rank $K$ has an eigen-decomposition (which can be efficiently computed):

$$M = \sum_{t=1}^{K} \lambda_t v_t v_t^T,$$

where $\{\lambda_t\}_{t=1}^{K}$ are eigenvalues, $\{v_t\}_{t=1}^{K}$ are the corresponding eigenvectors, which are orthonormal (i.e., orthogonal & unit length). Decomposition is unique iff $\{\lambda_t\}_{t=1}^{K}$ are distinct. (Up to sign of $v_t$s.) For (symmetric) tensors of order $p \geq 3$: an analogous decomposition is not guaranteed to exist.
Aside: eigenvalue decomposition

**Recall**: every symmetric matrix $M \in \mathbb{R}^{d \times d}$ of rank $K$ has an *eigen-decomposition* (which can be efficiently computed):

$$
M = \sum_{t=1}^{K} \lambda_t v_t v_t^T,
$$

- $\{\lambda_t\}_{t=1}^{K}$ are *eigenvalues*,
- $\{v_t\}_{t=1}^{K}$ are the corresponding *eigenvectors*, which are orthonormal (i.e., orthogonal & unit length).
- Decomposition is unique iff $\{\lambda_t\}_{t=1}^{K}$ are distinct. (Up to sign of $v_t$s.)
Aside: eigenvalue decomposition

Recall: every symmetric matrix $M \in \mathbb{R}^{d \times d}$ of rank $K$ has an eigen-decomposition (which can be efficiently computed):

$$M = \sum_{t=1}^{K} \lambda_t v_t v_t^T,$$

- $\{\lambda_t\}_{t=1}^{K}$ are eigenvalues,
- $\{v_t\}_{t=1}^{K}$ are the corresponding eigenvectors, which are orthonormal (i.e., orthogonal & unit length).
- Decomposition is unique iff $\{\lambda_t\}_{t=1}^{K}$ are distinct.
  (Up to sign of $v_t$s.)

For (symmetric) tensors of order $p \geq 3$:
- an analogous decomposition is **not** guaranteed to exist.
Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

\[ M = \sum_{t=1}^{K} v_t \otimes v_t, \]  
(e.g., word pairs)

\[ T = \sum_{t=1}^{K} \lambda_t \cdot v_t \otimes v_t \otimes v_t. \]  
(e.g., word triples)

Here, we assume \( \{v_t\}_{t=1}^{K} \) are linearly independent, and \( \{\lambda_t\}_{t=1}^{K} \) are positive.
Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

\[ M = \sum_{t=1}^{K} v_t \otimes v_t, \]  
(e.g., word pairs)

and

\[ T = \sum_{t=1}^{K} \lambda_t \cdot v_t \otimes v_t \otimes v_t. \]  
(e.g., word triples)

Here, we assume \( \{v_t\}_{t=1}^{K} \) are linearly independent, and \( \{\lambda_t\}_{t=1}^{K} \) are positive.

- \( M \) is positive semidefinite of rank \( K \).
Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

\[ M = \sum_{t=1}^{K} v_t \otimes v_t, \quad \text{(e.g., word pairs)} \]

and

\[ T = \sum_{t=1}^{K} \lambda_t \cdot v_t \otimes v_t \otimes v_t. \quad \text{(e.g., word triples)} \]

Here, we assume \( \{v_t\}_{t=1}^{K} \) are linearly independent, and \( \{\lambda_t\}_{t=1}^{K} \) are positive.

- \( M \) is positive semidefinite of rank \( K \).
- \( M \) determines inner product system on \( \text{span}\{v_t\}_{t=1}^{K} \) s.t. \( \{v_t\}_{t=1}^{K} \) are orthonormal:

\[ \langle x, y \rangle_M := x^\top M^\dagger y. \]
Reduction to orthonormal case

Suppose we have (estimates of) moments of the form

\[ M = \sum_{t=1}^{K} v_t \otimes v_t, \]  
(e.g., word pairs)

and

\[ T = \sum_{t=1}^{K} \lambda_t \cdot v_t \otimes v_t \otimes v_t. \]  
(e.g., word triples)

Here, we assume \( \{v_t\}_{t=1}^{K} \) are linearly independent, and \( \{\lambda_t\}_{t=1}^{K} \) are positive.

- \( M \) is positive semidefinite of rank \( K \).
- \( M \) determines inner product system on \( \text{span} \{v_t\}_{t=1}^{K} \) s.t. \( \{v_t\}_{t=1}^{K} \) are orthonormal:

  \[ \langle x, y \rangle_M := x^\top M^\dagger y. \]

- \( \therefore \) Can assume \( d = K \) and \( \{v_t\}_{t=1}^{d} \) are orthonormal.  
  (Similar to PCA; called “whitening” in signal processing context.)
Orthogonally decomposable tensors \((d = K)\)

**Goal:** Given tensor \(T = \sum_{t=1}^{d} \lambda_t \cdot v_t \otimes v_t \otimes v_t \in \mathbb{R}^{d \times d \times d}\) where:

- \(\{v_t\}_{t=1}^{d}\) are orthonormal;
- all \(\lambda_t > 0\);

approximately recover \(\{(v_t, \lambda_t)\}_{t=1}^{d}\).
Exact orthogonally decomposable tensor
(Zhang & Golub, 2001)

Matching moments:

\[
\{(\hat{v}_t, \hat{\lambda}_t)\}_{t=1}^d := \arg \min_{\{(x_t, \sigma_t)\}_{t=1}^d} \left\| T - \sum_{t=1}^d \sigma_t \cdot x_t \otimes x_t \otimes x_t \right\|_F^2.
\]  

(Here, \( \| \cdot \|_F \) is “Frobenius norm”, just like for matrices.)
Exact orthogonally decomposable tensor
(Zhang & Golub, 2001)

Matching moments:

\[
\left\{ (\hat{v}_t, \hat{\lambda}_t) \right\}_{t=1}^d := \arg \min_{\left\{ (x_t, \sigma_t) \right\}_{t=1}^d} \left\| T - \sum_{t=1}^d \sigma_t \cdot x_t \otimes x_t \otimes x_t \right\|_F^2.
\]

(Here, \( \| \cdot \|_F \) is “Frobenius norm”, just like for matrices.)

▶ Greedy approach:
  ▶ Find best rank-1 approximation:

\[
(\hat{v}, \hat{\lambda}) := \arg \min_{\|x\|=1, \sigma \geq 0} \| T - \sigma \cdot x \otimes x \otimes x \|_F^2.
\]

▶ “Deflate” \( T := T - \hat{\lambda} \cdot \hat{v} \otimes \hat{v} \otimes \hat{v} \) and repeat.
Exact orthogonally decomposable tensor
(Zhang & Golub, 2001)

Matching moments:

\[ \{(\hat{v}_t, \hat{\lambda}_t)\}_{t=1}^d := \arg \min_{\{(x_t, \sigma_t)\}_{t=1}^d} \left\| T - \sum_{t=1}^d \sigma_t \cdot x_t \otimes x_t \otimes x_t \right\|_F^2. \]

(Here, \(\| \cdot \|_F\) is “Frobenius norm”, just like for matrices.)

▶ Greedy approach:
  ▶ Find best rank-1 approximation:

\[ \hat{v} := \arg \max_{\|x\|=1} T(x, x, x), \quad \hat{\lambda} := T(\hat{v}, \hat{v}, \hat{v}). \]

▶ “Deflate” \(T := T - \hat{\lambda} \cdot \hat{v} \otimes \hat{v} \otimes \hat{v}\) and repeat.
Claim: Local maximizers of the function

\[ x \mapsto T(x, x, x) = \sum_{i,j,k} T_{i,j,k} \cdot x_i x_j x_k \]

(over the unit ball) are \( \{v_t\}_{t=1}^d \), and

\[ T(v_t, v_t, v_t) = \lambda_t, \quad t \in [d]. \]
Claim: Local maximizers of the function

\[ x \mapsto T(x, x, x) = \sum_{i,j,k} T_{i,j,k} \cdot x_i x_j x_k = \sum_{t=1}^{d} \lambda_t \cdot \langle v_t, x \rangle^3 \]

(over the unit ball) are \( \{v_t\}_{t=1}^{d} \), and

\[ T(v_t, v_t, v_t) = \lambda_t, \quad t \in [d]. \]
Claim: Local maximizers of the function

\[ x \mapsto T(x, x, x) = \sum_{i,j,k} T_{i,j,k} \cdot x_i x_j x_k = \sum_{t=1}^{d} \lambda_t \cdot \langle v_t, x \rangle^3 \]

(over the unit ball) are \( \{v_t\}_{t=1}^{d} \), and

\[ T(v_t, v_t, v_t) = \lambda_t, \quad t \in [d]. \]

Corollary: decomposition of \( T \) as \( \sum_{t=1}^{K} \lambda_t \cdot v_t^{\otimes 3} \) is unique!
Proof

By linearity and orthogonality:

\[
T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s^\otimes 3)(v_t, v_t, v_t)
\]
Proof

By linearity and orthogonality:

\[ T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s \otimes^3) (v_t, v_t, v_t) = \sum_{s=1}^{d} \left\{ \begin{array}{ll} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{array} \right. \]
Proof

By linearity and orthogonality:

\[ T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s \otimes v_s \otimes v_s)(v_t, v_t, v_t) = \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} = \lambda_t. \]
Proof

By linearity and orthogonality:

\[ T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s^{\otimes 3})(v_t, v_t, v_t) = \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} = \lambda_t. \]

WLOG assume \( v_t = e_t \), so optimization problem is

\[
\max_{x \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1.
\]
Proof

By linearity and orthogonality:

\[
T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s \otimes 3)(v_t, v_t, v_t) = \sum_{s=1}^{d} \begin{cases} 
\lambda_s & \text{if } s = t \\
0 & \text{if } s \neq t
\end{cases} = \lambda_t.
\]

WLOG assume \(v_t = e_t\), so optimization problem is

\[
\max_{x \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1.
\]

If both \(x_1\) and \(x_2\) are non-zero, then

\[
\lambda_1 x_1^3 + \lambda_2 x_2^3 < \lambda_1 x_1^2 + \lambda_2 x_2^2
\]
Proof

By linearity and orthogonality:

\[ T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s^\otimes 3)(v_t, v_t, v_t) = \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} = \lambda_t. \]

WLOG assume \( v_t = e_t \), so optimization problem is

\[
\max_{x \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1.
\]

If both \( x_1 \) and \( x_2 \) are non-zero, then

\[
\lambda_1 x_1^3 + \lambda_2 x_2^3 < \lambda_1 x_1^2 + \lambda_2 x_2^2 \leq \max\{\lambda_1, \lambda_2\}.
\]
Proof

By linearity and orthogonality:

$$T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s^\otimes 3)(v_t, v_t, v_t) = \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} = \lambda_t.$$  

WLOG assume $v_t = e_t$, so optimization problem is

$$\max_{x \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1.$$  

If both $x_1$ and $x_2$ are non-zero, then

$$\lambda_1 x_1^3 + \lambda_2 x_2^3 < \lambda_1 x_1^2 + \lambda_2 x_2^2 \leq \max\{\lambda_1, \lambda_2\}.$$  

So better to put all energy on a single coordinate.

$\therefore$ Local maximizers are $e_1, e_2, \ldots, e_d$.  

Proof

By linearity and orthogonality:

\[ T(v_t, v_t, v_t) = \sum_{s=1}^{d} (\lambda_s \cdot v_s^\otimes 3)(v_t, v_t, v_t) = \sum_{s=1}^{d} \begin{cases} \lambda_s & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases} = \lambda_t. \]

WLOG assume \( v_t = e_t \), so optimization problem is

\[
\max_{x \in \mathbb{R}^d} \sum_{t=1}^{d} \lambda_t x_t^3 \quad \text{s.t.} \quad \sum_{t=1}^{d} x_t^2 \leq 1.
\]

If both \( x_1 \) and \( x_2 \) are non-zero, then

\[
\lambda_1 x_1^3 + \lambda_2 x_2^3 < \lambda_1 x_1^2 + \lambda_2 x_2^2 \leq \max\{\lambda_1, \lambda_2\}.
\]

So better to put all energy on a single coordinate.

\[ \therefore \text{Local maximizers are } v_1, v_2, \ldots, v_d. \]
Uniqueness of orthogonal decompositions

What we have seen so far:

1. When components $\{v_t\}_{t=1}^d$ are linearly independent:
   - Reduce decomposition problem to *orthogonal tensor decomposition*, where components are orthonormal.
Uniqueness of orthogonal decompositions

What we have seen so far:

1. When components $\{v_t\}_{t=1}^d$ are linearly independent:
   - Reduce decomposition problem to orthogonal tensor decomposition, where components are orthonormal.

2. For orthogonally decomposable tensors $T$, local maximizers of the function

$$x \mapsto T(x, x, x)$$

(over the unit ball) are $\{v_t\}_{t=1}^d$. 
Uniqueness of orthogonal decompositions

What we have seen so far:

1. When components $\{v_t\}_{t=1}^d$ are linearly independent:
   - Reduce decomposition problem to *orthogonal tensor decomposition*, where components are orthonormal.

2. For orthogonally decomposable tensors $T$, local maximizers of the function
   $$ x \mapsto T(x, x, x) $$
   (over the unit ball) are $\{v_t\}_{t=1}^d$.

Algorithm: use gradient ascent to find all of the local maximizers, which are exactly $v_t$.

(Can use “deflation” to remove components from $T$ that you’ve already found.)
Application to topic model parameters

Probabilities of word triples as third-order tensor:

\[
T = \sum_{t=1}^{K} w_t P_t \otimes P_t \otimes P_t = \sum_{t=1}^{K} v_t \otimes v_t \otimes v_t
\]

for \( v_t = w_t^{1/3} P_t \).
Application to topic model parameters

Probabilities of word triples as third-order tensor:

\[ T = \sum_{t=1}^{K} w_t P_t \otimes P_t \otimes P_t = \sum_{t=1}^{K} v_t \otimes v_t \otimes v_t \]

for \( v_t = w_t^{1/3} P_t \).

- About linear independence condition on \( \{v_t\}_{t=1}^{K} \):

\[ \{v_t\}_{t=1}^{K} \text{ are linearly independent} \iff \{P_t\}_{t=1}^{K} \text{ are linearly independent and all } w_t > 0. \]
Application to topic model parameters

Probabilities of word triples as third-order tensor:

\[
T = \sum_{t=1}^{K} w_t P_t \otimes P_t \otimes P_t = \sum_{t=1}^{K} v_t \otimes v_t \otimes v_t
\]

for \( v_t = w_t^{1/3} P_t \).

- About linear independence condition on \( \{v_t\}_{t=1}^{K} \):

  \[
  \{v_t\}_{t=1}^{K} \text{ are linearly independent} \\
  \Leftrightarrow \{P_t\}_{t=1}^{K} \text{ are linearly independent and all } w_t > 0.
  \]

- Can recover \( \{P_t\}_{t=1}^{K} \) from \( \{c_t v_t\}_{t=1}^{K} \) for any \( c_t \neq 0 \).
Application to topic model parameters

Probabilities of word triples as third-order tensor:

\[ T = \sum_{t=1}^{K} w_t P_t \otimes P_t \otimes P_t = \sum_{t=1}^{K} v_t \otimes v_t \otimes v_t \]

for \[ v_t = w_t^{1/3} P_t. \]

- About linear independence condition on \( \{v_t\}_{t=1}^{K} \):
  \[ \{v_t\}_{t=1}^{K} \text{ are linearly independent} \iff \{P_t\}_{t=1}^{K} \text{ are linearly independent and all } w_t > 0. \]

- Can recover \( \{P_t\}_{t=1}^{K} \) from \( \{c_t v_t\}_{t=1}^{K} \) for any \( c_t \neq 0 \).

- Can recover \( \{(P_t, w_t)\}_{t=1}^{K} \) from \( \{P_t\}_{t=1}^{K} \) and \( T \).
Recap

- Parameters of topic model for single-topic documents (satisfying linear independence condition) can be efficiently recovered from distribution of three-word documents.
Recap

- Parameters of topic model for single-topic documents (satisfying linear independence condition) can be efficiently recovered from distribution of three-word documents.

- Two-word documents not sufficient (without further assumptions).
Recap

- Parameters of topic model for single-topic documents (satisfying linear independence condition) can be efficiently recovered from distribution of three-word documents.

- Two-word documents not sufficient (without further assumptions).

- Variational characterization of orthogonally decomposable tensors leads to simple and efficient algorithms!
Illustrative empirical results

- Corpus: 300,000 New York Times articles.
- Vocabulary size: 102,660 words.
- Set number of topics $K := 50$.

**Model predictive performance:**
$\approx 4\sim 8 \times$ speed-up over Gibbs sampling for LDA;

![Graph showing log loss vs. training time](image-url)
Illustrative empirical results

**Sample topics:** (showing top 10 words for each topic)

<table>
<thead>
<tr>
<th>Econ.</th>
<th>Baseball</th>
<th>Edu.</th>
<th>Health care</th>
<th>Golf</th>
</tr>
</thead>
<tbody>
<tr>
<td>sales</td>
<td>run</td>
<td>school</td>
<td>drug</td>
<td>player</td>
</tr>
<tr>
<td>economic</td>
<td>inning</td>
<td>student</td>
<td>patient</td>
<td>tiger_wood</td>
</tr>
<tr>
<td>consumer</td>
<td>hit</td>
<td>teacher</td>
<td>million</td>
<td>won</td>
</tr>
<tr>
<td>major</td>
<td>game</td>
<td>program</td>
<td>company</td>
<td>shot</td>
</tr>
<tr>
<td>home</td>
<td>season</td>
<td>official</td>
<td>doctor</td>
<td>play</td>
</tr>
<tr>
<td>indicator</td>
<td>home</td>
<td>public</td>
<td>companies</td>
<td>round</td>
</tr>
<tr>
<td>weekly</td>
<td>right</td>
<td>children</td>
<td>percent</td>
<td>win</td>
</tr>
<tr>
<td>order</td>
<td>games</td>
<td>high</td>
<td>cost</td>
<td>tournament</td>
</tr>
<tr>
<td>claim</td>
<td>dodger</td>
<td>education</td>
<td>program</td>
<td>tour</td>
</tr>
<tr>
<td>scheduled</td>
<td>left</td>
<td>district</td>
<td>health</td>
<td>right</td>
</tr>
</tbody>
</table>
### Sample topics: (showing top 10 words for each topic)

<table>
<thead>
<tr>
<th>Invest.</th>
<th>Election</th>
<th>auto race</th>
<th>Child’s Lit.</th>
<th>Afghan War</th>
</tr>
</thead>
<tbody>
<tr>
<td>percent</td>
<td>al_gore</td>
<td>car</td>
<td>book</td>
<td>taliban</td>
</tr>
<tr>
<td>stock</td>
<td>campaign</td>
<td>race</td>
<td>children</td>
<td>attack</td>
</tr>
<tr>
<td>market</td>
<td>president</td>
<td>driver</td>
<td>ages</td>
<td>afghanistan</td>
</tr>
<tr>
<td>fund</td>
<td>george_bush</td>
<td>team</td>
<td>author</td>
<td>official</td>
</tr>
<tr>
<td>investor</td>
<td>bush</td>
<td>won</td>
<td>read</td>
<td>military</td>
</tr>
<tr>
<td>companies</td>
<td>clinton</td>
<td>win</td>
<td>newspaper</td>
<td>u_s</td>
</tr>
<tr>
<td>analyst</td>
<td>vice</td>
<td>racing</td>
<td>web</td>
<td>united_states</td>
</tr>
<tr>
<td>money</td>
<td>presidential</td>
<td>track</td>
<td>writer</td>
<td>terrorist</td>
</tr>
<tr>
<td>investment</td>
<td>million</td>
<td>season</td>
<td>written</td>
<td>war</td>
</tr>
<tr>
<td>economy</td>
<td>democratic</td>
<td>lap</td>
<td>sales</td>
<td>bin</td>
</tr>
</tbody>
</table>
Illustrative empirical results

Sample topics: (showing top 10 words for each topic)

<table>
<thead>
<tr>
<th>Web</th>
<th>Antitrust</th>
<th>TV</th>
<th>Movies</th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td>com</td>
<td>court</td>
<td>show</td>
<td>film</td>
<td>music</td>
</tr>
<tr>
<td>www</td>
<td>case</td>
<td>network</td>
<td>movie</td>
<td>song</td>
</tr>
<tr>
<td>site</td>
<td>law</td>
<td>season</td>
<td>director</td>
<td>group</td>
</tr>
<tr>
<td>web</td>
<td>lawyer</td>
<td>nbc</td>
<td>play</td>
<td>part</td>
</tr>
<tr>
<td>sites</td>
<td>federal</td>
<td>cb</td>
<td>character</td>
<td>new_york</td>
</tr>
<tr>
<td>information</td>
<td>government</td>
<td>program</td>
<td>actor</td>
<td>company</td>
</tr>
<tr>
<td>online</td>
<td>decision</td>
<td>television</td>
<td>show</td>
<td>million</td>
</tr>
<tr>
<td>mail</td>
<td>trial</td>
<td>series</td>
<td>movies</td>
<td>band</td>
</tr>
<tr>
<td>internet</td>
<td>microsoft</td>
<td>night</td>
<td>million</td>
<td>show</td>
</tr>
<tr>
<td>telegram</td>
<td>right</td>
<td>new_york</td>
<td>part</td>
<td>album</td>
</tr>
</tbody>
</table>

etc.
**Computation**

*Caveat*: forming and computing with a third-order tensor $T$ generally requires cubic space.
Computation

**Caveat:** forming and computing with a third-order tensor $T$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $T$: e.g.,

$$\hat{T} = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i,$$

where $\text{data}_i$ is a tensor involving only the $i$-th data point.
Computation

**Caveat**: forming and computing with a third-order tensor $T$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $T$: e.g.,

$$\hat{T} = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i,$$

where $\text{data}_i$ is a tensor involving only the $i$-th data point.

- Our algorithms will only involve $\hat{T}$ through *evaluations* of $\hat{T}$ at (several) given arguments, say, $x, y, z$. 
Computation

**Caveat**: forming and computing with a third-order tensor $T$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $T$: e.g.,

\[
\hat{T} = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i ,
\]

where $\text{data}_i$ is a tensor involving only the $i$-th data point.

- Our algorithms will only involve $\hat{T}$ through evaluations of $\hat{T}$ at (several) given arguments, say, $x, y, z$.

By linearity:

\[
\hat{T}(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i(x, y, z) .
\]


**Computation**

**Caveat**: forming and computing with a third-order tensor $T$ generally requires cubic space.

- Fortunately, the tensor we often work with is an *empirical estimate* of a $T$: e.g.,

$$\hat{T} = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i ,$$

where $\text{data}_i$ is a tensor involving only the $i$-th data point.

- Our algorithms will only involve $\hat{T}$ through *evaluations* of $\hat{T}$ at (several) given arguments, say, $x, y, z$.

By linearity:

$$\hat{T}(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i(x, y, z) .$$

- Often: $\text{data}_i(x, y, z)$ is easy to compute, even without forming any tensors!
Caveat: forming and computing with a third-order tensor $T$ generally requires cubic space.

- Fortunately, the tensor we often work with is an empirical estimate of a $T$: e.g.,

$$\hat{T} = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i,$$

where $\text{data}_i$ is a tensor involving only the $i$-th data point.

- Our algorithms will only involve $\hat{T}$ through evaluations of $\hat{T}$ at (several) given arguments, say, $x, y, z$.

By linearity:

$$\hat{T}(x, y, z) = \frac{1}{n} \sum_{i=1}^{n} \text{data}_i(x, y, z).$$

- Often: $\text{data}_i(x, y, z)$ is easy to compute, even without forming any tensors! $\longrightarrow$ Linear time/space algorithms.
Learning algorithms

- Estimation via **method-of-moments**:
  
  1. *Estimate* distribution of three-word documents $\to \hat{T}$
     
     *(empirical moment tensor)*.
  
  2. *Approximately decompose* $\hat{T}$ $\to$ estimates $\{(\hat{P}_t, \hat{w}_t)\}_{t=1}^K$. 

Issues:

1. Accuracy of moment estimates?
   - Can more reliably estimate lower-order moments; distribution-specific sample complexity bounds.

2. Robustness of (approximate) tensor decomposition?
   - In some sense, more stable than matrix eigen-decomposition (Mu, H., & Goldfarb, 2015).

3. Generality beyond simple topic models?
   - Next: Moment decompositions for other models.
Learning algorithms

- Estimation via **method-of-moments**:
  1. *Estimate* distribution of three-word documents $\rightarrow \hat{T}$ (*empirical moment tensor*).
  2. *Approximately decompose* $\hat{T} \rightarrow$ estimates $\{(\hat{P}_t, \hat{w}_t)\}_{t=1}^K$.

- **Issues**:
  1. Accuracy of *moment estimates*?
  2. Robustness of (*approximate*) *tensor decomposition*?
  3. *Generality* beyond simple topic models?
Learning algorithms

- Estimation via method-of-moments:
  1. Estimate distribution of three-word documents → $\hat{T}$ (empirical moment tensor).
  2. Approximately decompose $\hat{T}$ → estimates $\{(\hat{P}_t, \hat{w}_t)\}_{t=1}^K$.

- Issues:
  1. Accuracy of moment estimates?
     Can more reliably estimate lower-order moments; distribution-specific sample complexity bounds.
  2. Robustness of (approximate) tensor decomposition?
     In some sense, more stable than matrix eigen-decomposition (Mu, H., & Goldfarb, 2015)!
  3. Generality beyond simple topic models?
Learning algorithms

▶ Estimation via method-of-moments:

1. Estimate distribution of three-word documents $\rightarrow \hat{T}$ (empirical moment tensor).

2. Approximately decompose $\hat{T} \rightarrow$ estimates $\{(\hat{P}_t, \hat{w}_t)\}_{t=1}^K$.

▶ Issues:

1. Accuracy of moment estimates?
   Can more reliably estimate lower-order moments; distribution-specific sample complexity bounds.

2. Robustness of (approximate) tensor decomposition?
   In some sense, more stable than matrix eigen-decomposition (Mu, H., & Goldfarb, 2015)!

3. Generality beyond simple topic models?

Next: Moment decompositions for other models.
2. Moment decompositions for other models
Moment decompositions

Some examples of usable moment decompositions.

1. Two classical mixture models.
Mixture model #1: Mixtures of spherical Gaussians

\[ H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \text{ (hidden)}; \]
\[ X \mid H = t \sim \text{Normal}(\mu_t, \sigma_t^2 I_d), \quad t \in [K]. \]
Mixture model #1: Mixtures of spherical Gaussians

\[ H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \text{ (hidden)}; \]
\[ X \mid H = t \sim \text{Normal}(\mu_t, \sigma^2 I_d), \quad t \in [K]. \]

(For simplicity, restrict \( \sigma_1 = \sigma_2 = \cdots = \sigma_K = \sigma \).)
Mixture model #1: Mixtures of spherical Gaussians

\[ H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \text{ (hidden)}; \]
\[ X \mid H = t \sim \text{Normal}(\mu_t, \sigma^2 I_d), \quad t \in [K]. \]

(For simplicity, restrict \( \sigma_1 = \sigma_2 = \cdots = \sigma_K = \sigma \).)

Generative process:

\[ X = Y + \sigma Z \]

where \( \Pr(Y = \mu_t) = \pi_t \), and \( Z \sim \text{Normal}(0, I_d) \) (indep. of \( Y \)).
Using moments for spherical Gaussian mixtures

We’ll see two ways to use low-order moments.
Using moments for spherical Gaussian mixtures

We’ll see two ways to use low-order moments.

First- and second-order moments:

\[ \mathbb{E}(X) \in \mathbb{R}^d \quad \text{and} \quad \mathbb{E}(X \otimes X) \in \mathbb{R}^{d \times d}. \]
Using moments for spherical Gaussian mixtures

We’ll see two ways to use low-order moments.

**First- and second-order moments:**

\[
\mathbb{E}(X) \in \mathbb{R}^d \quad \text{and} \quad \mathbb{E}(X \otimes X) \in \mathbb{R}^{d \times d}.
\]

**Claim** (Vempala & Wang, 2002):

Span of top \( K \) eigenvectors of \( \mathbb{E}(X \otimes X) \) contains \( \{\mu_t\}_{t=1}^K \).

(\( K \)-dimensional Principal Component Analysis (PCA) subspace.)
Proof

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.
Proof

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- $K = 1$ (just a single Gaussian):
  
  What is the 1-dimensional PCA subspace?
Proof

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?

\[
\mathbb{E}(X \otimes X) = \mu_1 \otimes \mu_1 + \sigma^2 I_d.
\]
Proof

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?

\[
\mathbb{E}(X \otimes X) = \mu_1 \otimes \mu_1 + \sigma^2 I_d.
\]

Variance in direction $\mathbf{v}$ (with $\|\mathbf{v}\| = 1$):

\[
\mathbf{v}^\top \mathbb{E}(X \otimes X) \mathbf{v}
\]
Proof

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- $K = 1$ (just a single Gaussian):
  
  What is the 1-dimensional PCA subspace?

\[
\mathbb{E}(X \otimes X) = \mu_1 \otimes \mu_1 + \sigma^2 I_d.
\]

Variance in direction $\nu$ (with $\|\nu\| = 1$):

\[
\nu^\top \mathbb{E}(X \otimes X) \nu = \nu^\top (\mu_1 \otimes \mu_1 + \sigma^2 I_d) \nu
\]
Proof

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the 1-dimensional PCA subspace?

\[ \mathbb{E}(X \otimes X) = \mu_1 \otimes \mu_1 + \sigma^2 I_d. \]

Variance in direction $\mathbf{v}$ (with $\|\mathbf{v}\| = 1$):

\[ \mathbf{v}^\top \mathbb{E}(X \otimes X) \mathbf{v} = \mathbf{v}^\top (\mu_1 \otimes \mu_1 + \sigma^2 I_d) \mathbf{v} = (\mathbf{v}^\top \mu_1)^2 + \sigma^2. \]
Proof

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- $K = 1$ (just a single Gaussian):
  
  What is the 1-dimensional PCA subspace?

\[ \mathbb{E}(X \otimes X) = \mu_1 \otimes \mu_1 + \sigma^2 I_d. \]

Variance in direction $\nu$ (with $\|\nu\| = 1$):

\[ \nu^\top \mathbb{E}(X \otimes X) \nu = \nu^\top (\mu_1 \otimes \mu_1 + \sigma^2 I_d) \nu = (\nu^\top \mu_1)^2 + \sigma^2. \]

**Best direction** (1-dim. PCA subspace): $\nu = \pm \mu_1/\|\mu_1\|$. 
Proof (continued)

**Key fact:** \( k \)-dimensional PCA subspace (based on \( \mathbb{E}(X \otimes X) \)) captures as much of overall variance as any other \( k \)-dim. subspace.

- \( K = 1 \) (just a single Gaussian):
  
  What is the \( k \)-dimensional PCA subspace?

\[ \mathbb{R}^d \]

\[ \mu_1 \]

origin
Proof (continued)

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- $K = 1$ (just a single Gaussian):
  What is the $k$-dimensional PCA subspace?

**Answer:** any $k$-dim. subspace containing $\mu_1$. 

![Diagram](image-url)
Proof (continued)

**Key fact:** \(k\)-dimensional PCA subspace (based on \(E(X \otimes X)\)) captures as much of overall variance as any other \(k\)-dim. subspace.

- General \(K\) (mixture of \(K\) Gaussians):
  
  What is the \(K\)-dimensional PCA subspace?
Proof (continued)

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- General $K$ (mixture of $K$ Gaussians):
  What is the $K$-dimensional PCA subspace?

![Diagram of three Gaussian distributions with means $\mu_1$, $\mu_2$, and $\mu_3$ in $\mathbb{R}^d$.]

\[
\mathbb{E}(X \otimes X) = \sum_{t=1}^{K} \pi_t \cdot \mu_t \otimes \mu_t + \sigma^2 I_d.
\]
Proof (continued)

**Key fact**: $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- General $K$ (mixture of $K$ Gaussians):
  What is the $K$-dimensional PCA subspace?

\[
\mathbb{E}(X \otimes X) = \sum_{t=1}^{K} \pi_t \cdot \mu_t \otimes \mu_t + \sigma^2 I_d.
\]

**Answer**: any $K$-dim. subspace containing $\mu_1, \ldots, \mu_K$. □
Proof (continued)

**Key fact:** $k$-dimensional PCA subspace (based on $\mathbb{E}(X \otimes X)$) captures as much of overall variance as any other $k$-dim. subspace.

- **General $K$ (mixture of $K$ Gaussians):**
  What is the $K$-dimensional PCA subspace?

\[
\mathbb{E}(X \otimes X) = \sum_{t=1}^{K} \pi_t \cdot \mu_t \otimes \mu_t + \sigma^2 I_d.
\]

**Answer:** any $K$-dim. subspace containing $\mu_1, \ldots, \mu_K$.

How does this help with learning mixtures of Gaussians?
Use of moments for mixtures of spherical Gaussians

**Separation** *(Dasgupta, 1999):*

# standard deviations between component means

\[
\text{sep} := \min_{i \neq j} \frac{\|\mu_i - \mu_j\|}{\sigma}.
\]
Use of moments for mixtures of spherical Gaussians

**Separation** (Dasgupta, 1999):

\# standard deviations between component means

\[
\text{sep} := \min_{i \neq j} \frac{\| \mu_i - \mu_j \|}{\sigma}.
\]

▶ (Dasgupta & Schulman, 2000):

Distance-based clustering (e.g., EM) works when \( \text{sep} \gtrsim d^{1/4} \).

(Vempala & Wang, 2002):

Problem becomes \( K \)-dimensional via PCA (assume \( K \leq d \)).

Required separation reduced to \( \text{sep} \gtrsim K^{1/4} \).

Third-order moments identify the mixture distribution when \( \{ \mu_t \}_{t=1}^K \) are lin. indpt.; sep may be arbitrarily close to zero.

(Belkin & Sinha, 2010; Moitra & Valiant, 2010):

General Gaussians & no minimum sep, but \( K \)-th order moments.
Use of moments for mixtures of spherical Gaussians

**Separation** (Dasgupta, 1999):  
# standard deviations between component means

\[
\text{sep} := \min_{i \neq j} \frac{\| \mu_i - \mu_j \|}{\sigma}.
\]

▶ (Dasgupta & Schulman, 2000):  
Distance-based clustering (e.g., EM) works when \( \text{sep} \gtrsim d^{1/4} \).

▶ (Vempala & Wang, 2002):  
Problem becomes \( K \)-dimensional via PCA (assume \( K \leq d \)).  
Required separation reduced to \( \text{sep} \gtrsim K^{1/4} \).
Use of moments for mixtures of spherical Gaussians

**Separation** (Dasgupta, 1999):

# standard deviations between component means

\[
sep := \min_{i \neq j} \frac{\| \mu_i - \mu_j \|}{\sigma}.
\]

▶ (Dasgupta & Schulman, 2000):
Distance-based clustering (e.g., EM) works when \( sep \gtrsim d^{1/4} \).

▶ (Vempala & Wang, 2002):
Problem becomes \( K \)-dimensional via PCA (assume \( K \leq d \)).
Required separation reduced to \( sep \gtrsim K^{1/4} \).

**Third-order moments** identify the mixture distribution when \( \{ \mu_t \}_{t=1}^{K} \) are lin. indpt.; \( sep \) may be arbitrarily close to zero.
Use of moments for mixtures of spherical Gaussians

**Separation** (Dasgupta, 1999):

\# standard deviations between component means

\[
\text{sep} := \min_{i \neq j} \frac{\|\mu_i - \mu_j\|}{\sigma}.
\]

- (Dasgupta & Schulman, 2000):
  Distance-based clustering (e.g., EM) works when \(\text{sep} \gtrsim d^{1/4}\).

- (Vempala & Wang, 2002):
  Problem becomes \(K\)-dimensional via PCA (assume \(K \leq d\)).
  Required separation reduced to \(\text{sep} \gtrsim K^{1/4}\).

---

**Third-order moments** identify the mixture distribution when \(\{\mu_t\}_{t=1}^K\) are lin. indpt.; \(\text{sep}\) may be arbitrarily close to zero.

(Belkin & Sinha, 2010; Moitra & Valiant, 2010):
General Gaussians \& no minimum \(\text{sep}\), but \(K\)th-order moments.
Third-order moments of spherical Gaussian mixtures

Generative process:

\[ X = Y + \sigma Z \]

where \( \Pr(Y = \mu_t) = \pi_t \), and \( Z \sim \text{Normal}(0, I_d) \), \( Y \perp \perp Z \).

Third-order moment tensor:

\[ \mathbb{E}(X^{\otimes 3}) = \mathbb{E}\left(\{Y + \sigma Z\}^{\otimes 3}\right) \]
Third-order moments of spherical Gaussian mixtures

**Generative process:**

\[ X = Y + \sigma Z \]

where \( \Pr(Y = \mu_t) = \pi_t \), and \( Z \sim \text{Normal}(0, I_d) \), \( Y \perp\!\!\!\perp Z \).

Third-order moment tensor:

\[
\mathbb{E} \left( X \otimes^3 \right) = \mathbb{E} \left( \{Y + \sigma Z\} \otimes^3 \right) \\
= \mathbb{E} \left( Y \otimes^3 \right) + \sigma^2 \mathbb{E} \left( Y \otimes Z \otimes Z + Z \otimes Y \otimes Z + Z \otimes Z \otimes Y \right)
\]

(Above, \( \mu = \mathbb{E}(X) \) and \( \tau(\mu) \) is a third-order tensor involving only \( \mu \).)

Exercise: find explicit formula for \( \tau(\mu) \).
Third-order moments of spherical Gaussian mixtures

**Generative process:**

\[ X = Y + \sigma Z \]

where \( \Pr(Y = \mu_t) = \pi_t \), and \( Z \sim \text{Normal}(0, I_d) \), \( Y \perp \perp Z \).

Third-order moment tensor:

\[
\mathbb{E} \left( X \otimes^3 \right) = \mathbb{E} \left( \{ Y + \sigma Z \} \otimes^3 \right) \\
= \mathbb{E} \left( Y \otimes^3 \right) + \sigma^2 \mathbb{E} \left( Y \otimes Z \otimes Z + Z \otimes Y \otimes Z + Z \otimes Z \otimes Y \right) \\
= \sum_{t=1}^{K} \pi_t \cdot \mu_t \otimes^3 + \sigma^2 \tau(\mu). 
\]

(Above, \( \mu = \mathbb{E}(X) \) and \( \tau(\mu) \) is a third-order tensor involving only \( \mu \).)
Third-order moments of spherical Gaussian mixtures

Generative process:

\[ X = Y + \sigma Z \]

where \( \Pr(Y = \mu_t) = \pi_t \), and \( Z \sim \text{Normal}(0, I_d) \), \( Y \perp \perp Z \).

Third-order moment tensor:

\[
\mathbb{E} \left( X^{\otimes 3} \right) = \mathbb{E} \left( \{ Y + \sigma Z \}^{\otimes 3} \right) \\
= \mathbb{E} \left( Y^{\otimes 3} \right) + \sigma^2 \mathbb{E} \left( Y \otimes Z \otimes Z + Z \otimes Y \otimes Z + Z \otimes Z \otimes Y \right) \\
= \sum_{t=1}^{K} \pi_t \cdot \mu_t^{\otimes 3} + \sigma^2 \tau(\mu) .
\]

(Above, \( \mu = \mathbb{E}(X) \) and \( \tau(\mu) \) is a third-order tensor involving only \( \mu \).)

Exercise: find explicit formula for \( \tau(\mu) \).
Tensor decomposition for spherical Gaussian mixtures
(H. & Kakade, 2013)

\[ \mathbb{E}(X \otimes^3) = \sum_{t=1}^{K} \pi_t \cdot \mu_t^{\otimes 3} + \sigma^2 \tau(\mu). \]
Tensor decomposition for spherical Gaussian mixtures
(H. & Kakade, 2013)

\[ \mathbb{E} (X \otimes^3) = \sum_{t=1}^{K} \pi_t \cdot \mu_t \otimes^3 + \sigma^2 \tau(\mu). \]

Claim: \( \mu \) & \( \sigma^2 \) are simple functions of \( \mathbb{E}(X) \) & \( \mathbb{E}(X \otimes X) \).
Tensor decomposition for spherical Gaussian mixtures
(H. & Kakade, 2013)

\[
\mathbb{E}\left( X \otimes^3 \right) = \sum_{t=1}^{K} \pi_t \cdot \mu_t \otimes^3 + \sigma^2 \tau(\mu).
\]

**Claim:** \( \mu \) & \( \sigma^2 \) are simple functions of \( \mathbb{E}(X) \) & \( \mathbb{E}(X \otimes X) \).

**Claim:** If \( \{ \mu_t \}_{t=1}^{K} \) are linearly independent and all \( \pi_t > 0 \), then \( \{(\mu_t, \pi_t)\}_{t=1}^{K} \) are identifiable from

\[
T := \mathbb{E}(X \otimes^3) - \sigma^2 \tau(\mu) = \sum_{t=1}^{K} \pi_t \cdot \mu_t \otimes^3.
\]
Tensor decomposition for spherical Gaussian mixtures
(H. & Kakade, 2013)

\[
\mathbb{E}\left( X^{\otimes 3} \right) = \sum_{t=1}^{K} \pi_t \cdot \mu_t^{\otimes 3} + \sigma^2 \tau(\mu).
\]

**Claim:** \( \mu \) & \( \sigma^2 \) are simple functions of \( \mathbb{E}(X) \) & \( \mathbb{E}(X \otimes X) \).

**Claim:** If \( \{\mu_t\}_{t=1}^{K} \) are linearly independent and all \( \pi_t > 0 \), then \( \{(\mu_t, \pi_t)\}_{t=1}^{K} \) are identifiable from

\[
T := \mathbb{E}(X^{\otimes 3}) - \sigma^2 \tau(\mu) = \sum_{t=1}^{K} \pi_t \cdot \mu_t^{\otimes 3}.
\]

Can use tensor decomposition to recover \( \{(\mu_t, \pi_t)\}_{t=1}^{K} \) from \( T \).
Even more Gaussian mixtures

**Note**: Linear independence condition on \( \{\mu_t\}_{t=1}^K \) requires \( K \leq d \).
Even more Gaussian mixtures

Note: Linear independence condition on $\{\mu_t\}_{t=1}^K$ requires $K \leq d$.

- (Anderson, Belkin, Goyal, Rademacher, & Voss, 2014),
  (Bhaskara, Charikar, Moitra, & Vijayaraghavan, 2014)

Mixtures of $d^{O(1)}$ Gaussians (w/ simple or known covariance) via smoothed analysis and $O(1)$-order moments.
Even more Gaussian mixtures

**Note:** Linear independence condition on $\{\mu_t\}_{t=1}^K$ requires $K \leq d$.

- (Anderson, Belkin, Goyal, Rademacher, & Voss, 2014),
  (Bhaskara, Charikar, Moitra, & Vijayaraghavan, 2014)
Mixtures of $d^{O(1)}$ Gaussians (w/ simple or known covariance) via *smoothed analysis* and $O(1)$-order moments.

- (Ge, Huang, & Kakade, 2015)
Also with *unknown covariances of arbitrary shape*. 
Mixture model #2: Mixtures of linear regressions

\[
H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \text{ (hidden)};
\]

\[
X \sim \text{Normal}(\mu, \Sigma);
\]

\[
Y \mid H = t, X = x \sim \text{Normal}(\langle \beta_t, x \rangle, \sigma^2).
\]
Mixture model #2: Mixtures of linear regressions

\[ H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \text{ (hidden)}; \]
\[ X \sim \text{Normal}(\mu, \Sigma); \]
\[ Y \mid H = t, X = x \sim \text{Normal}(\langle \beta_t, x \rangle, \sigma^2). \]
Mixture model #2: Mixtures of linear regressions

\[ H \sim \text{Categorical}(\pi_1, \pi_2, \ldots, \pi_K) \quad (\text{hidden}) ; \]
\[ X \sim \text{Normal}(\mu, \Sigma) ; \]
\[ Y \mid H = t, X = x \sim \text{Normal}(\langle \beta_t, x \rangle, \sigma^2) . \]
Use of moments for mixtures of linear regressions

Second-order moments (assume $X \sim \text{Normal}(0, I_d)$):

$$
\mathbb{E}(Y^2 XX^\top) = 2 \sum_{t=1}^{K} \pi_t \cdot \beta_t \beta_t^\top
+ \left( \sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\beta_t\|^2 \right) I_d.
$$
Use of moments for mixtures of linear regressions

**Second-order moments** (assume $X \sim \text{Normal}(0, I_d)$):

$$
\mathbb{E}(Y^2 XX^\top) = 2 \sum_{t=1}^{K} \pi_t \cdot \beta_t \beta_t^\top + \left( \sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\beta_t\|^2 \right) I_d.
$$

- Span of top $K$ eigenvectors of $\mathbb{E}(Y^2 XX^\top)$ contains $\{\beta_t\}_{t=1}^{K}$. 

Using Stein's identity (1973), similar approach works for GLMs (Sun, Ioannidis, & Montanari, 2013).

Tensor decomposition approach: Can recover parameters $\{(\beta_t, \pi_t)\}_{t=1}^{K}$ with higher-order moments (Chaganty & Liang, 2013; Yi, Caramanis, & Sanghavi, 2014, 2016).

Also for GLMs, via Stein's identity (Sedghi & Anandkumar, 2014).
Use of moments for mixtures of linear regressions

**Second-order moments** (assume $X \sim \text{Normal}(0, I_d)$):

$$
\mathbb{E}(Y^2XX^\top) = 2 \sum_{t=1}^{K} \pi_t \cdot \beta_t \beta_t^\top + \left( \sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\beta_t\|^2 \right) I_d.
$$

- Span of top $K$ eigenvectors of $\mathbb{E}(Y^2XX^\top)$ contains $\{\beta_t\}_{t=1}^{K}$.
- Using Stein’s identity (1973), similar approach works for GLMs (Sun, Ioannidis, & Montanari, 2013).
Use of moments for mixtures of linear regressions

**Second-order moments** (assume \( \mathbf{X} \sim \text{Normal}(0, \mathbf{I}_d) \)):

\[
\mathbb{E}(Y^2 \mathbf{X} \mathbf{X}^\top) = 2 \sum_{t=1}^{K} \pi_t \cdot \beta_t \beta_t^\top + \left( \sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \|\beta_t\|^2 \right) \mathbf{I}_d.
\]

- Span of top \( K \) eigenvectors of \( \mathbb{E}(Y^2 \mathbf{X} \mathbf{X}^\top) \) contains \( \{\beta_t\}_{t=1}^{K} \).
- Using **Stein’s identity** (1973), similar approach works for GLMs (Sun, Ioannidis, & Montanari, 2013).

**Tensor decomposition approach:**
Can recover parameters \( \{(\beta_t, \pi_t)\}_{t=1}^{K} \) with higher-order moments (Chaganty & Liang, 2013; Yi, Caramanis, & Sanghavi, 2014, 2016).
Use of moments for mixtures of linear regressions

**Second-order moments** (assume $X \sim \text{Normal}(0, I_d)$):

$$
\mathbb{E}(Y^2XX^\top) = 2 \sum_{t=1}^{K} \pi_t \cdot \beta_t \beta_t^\top + \left( \sigma^2 + \sum_{t=1}^{K} \pi_t \cdot \| \beta_t \|^2 \right) I_d .
$$

- Span of top $K$ eigenvectors of $\mathbb{E}(Y^2XX^\top)$ contains $\{ \beta_t \}_{t=1}^{K}$.
- Using Stein’s identity (1973), similar approach works for GLMs (Sun, Ioannidis, & Montanari, 2013).

**Tensor decomposition approach:**
Can recover parameters $\{ (\beta_t, \pi_t) \}_{t=1}^{K}$ with higher-order moments (Chaganty & Liang, 2013; Yi, Caramanis, & Sanghavi, 2014, 2016).

Also for GLMs, via Stein’s identity (Sedghi & Anandkumar, 2014).
Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.
Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.

- Exploit distributional properties to determine usable moments.
Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.

- Exploit distributional properties to determine usable moments.

- Smoothed analysis: avoid linear independence condition for “most” mixture distributions.
Recap: mixtures of Gaussians and linear regressions

- Parameters of Gaussian mixture models and related models (satisfying linear independence condition) can be efficiently recovered from $O(1)$-order moments.

- Exploit distributional properties to determine usable moments.

- *Smoothed analysis*: avoid linear independence condition for “most” mixture distributions.

**Next**: Multi-view approach to finding usable moments.
Multi-view interpretation of topic model

**Recall:** Topic model for single-topic documents

- **$H$** topics (dists. over words) $\{P_t\}_{t=1}^K$.  
- Pick topic $H = t$ with prob. $w_t$ (hidden).  
- Word tokens $X_1, X_2, \ldots, X_L \overset{iid}{\sim} P_H$.  

![Diagram](image)
Multi-view interpretation of topic model

**Recall:** Topic model for single-topic documents

$K$ topics (dists. over words) $\{P_t\}_{t=1}^K$.

Pick topic $H = t$ with prob. $w_t$ (hidden).

Word tokens $X_1, X_2, \ldots, X_L \overset{iid}{\sim} P_H$.

**Key property:**

$X_1, X_2, \ldots, X_L$ conditionally independent given $H$. 

Some previous analyses:

- (Chaudhuri, Kakade, Livescu, & Sridharan, 2009) Multi-view Gaussian mixture models.
Multi-view interpretation of topic model

**Recall:** Topic model for single-topic documents

- **H**
- **X_1**, **X_2**, ..., **X_L**

\[ K \text{ topics (dists. over words)} \{P_t\}_{t=1}^{K}. \]

Pick topic **H = t** with prob. \( w_t \) (hidden).

Word tokens \( X_1, X_2, \ldots, X_L \) \( \text{iid} \sim P_H. \)

**Key property:**
\( X_1, X_2, \ldots, X_L \) conditionally independent given \( H \).

Each word token **X_i** provides new “view” of hidden variable **H**.
Multi-view interpretation of topic model

**Recall:** Topic model for single-topic documents

![Diagram of topic model](image)

- $K$ topics (dists. over words) $\{P_t\}_{t=1}^K$.
- Pick topic $H = t$ with prob. $w_t$ (hidden).
- Word tokens $X_1, X_2, \ldots, X_L \sim P_H$.

**Key property:**

$X_1, X_2, \ldots, X_L$ conditionally independent given $H$.

Each word token $X_i$ provides new “view” of hidden variable $H$.

**Some previous analyses:**

- (Blum & Mitchell, 1998)
  *Co-training* in semi-supervised learning.

- (Chaudhuri, Kakade, Livescu, & Sridharan, 2009)
  Multi-view Gaussian mixture models.
Multi-view mixture model

View 1: $X_1$  
View 2: $X_2$  
View 3: $X_3$
Multi-view mixture model

View 1: $X_1$  View 2: $X_2$  View 3: $X_3$
Multi-view mixture model

\[
E \left( X_1 \otimes X_2 \otimes X_3 \right) = \sum_{t=1}^{K} \pi_t \cdot \mu_t^{(1)} \otimes \mu_t^{(2)} \otimes \mu_t^{(3)}
\]

where \( \mu_t^{(i)} = E[X_i | H = t] \),

\( \pi_t = Pr(H = t) \).
Multi-view mixture model

\[
\mathbb{E} (X_1 \otimes X_2 \otimes X_3) = \sum_{t=1}^{K} \pi_t \cdot \mu_t^{(1)} \otimes \mu_t^{(2)} \otimes \mu_t^{(3)}
\]

where \( \mu_t^{(i)} = \mathbb{E}[X_i | H = t] \),

\( \pi_t = \text{Pr}(H = t) \).

**Tensor decomposition approach** works in this asymmetric case as long as \( \{\mu_t^{(j)}\}_{t=1}^{K} \) are lin. indpt. for each \( j \), and all \( \pi_t > 0 \).
Examples of multi-view mixture models
(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)
Examples of multi-view mixture models
(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

2. Hidden Markov models.

\[
H_1 \rightarrow H_2 \rightarrow H_3 \\
X_1 \rightarrow X_2 \rightarrow X_3
\]

\[
\rightarrow \\
H_2 \\
X_1 \rightarrow X_2 \rightarrow X_3
\]
Examples of multi-view mixture models
(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions. 
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

2. Hidden Markov models.

3. Phylogenetic trees.
   - $X_1, X_2, X_3$: genes of three extant species.
   - $H$: LCA of most closely related pair of species.
Examples of multi-view mixture models

(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

2. Hidden Markov models.

3. Phylogenetic trees.
   - $X_1, X_2, X_3$: genes of three extant species.
   - $H$: LCA of most closely related pair of species.

4. ...
Examples of multi-view mixture models

(Mossel & Roch, 2006; Anandkumar, H., & Kakade, 2012)

1. Mixtures of high-dimensional product distributions.  
   (E.g., mixtures of axis-aligned Gaussians, other topic models.)

2. Hidden Markov models.

   ![Hidden Markov Model Diagram]

3. Phylogenetic trees.
   
   ▶ $X_1, X_2, X_3$: genes of three extant species.
   
   ▶ $H$: LCA of most closely related pair of species.

4. ... 

Next: Single index models.
Single-index models

\[ X \sim \text{Normal}(0, I); \]
\[ Y \mid X = x \sim \text{Normal}(g(\langle \beta, x \rangle), \sigma^2). \]

Here, \( g: \mathbb{R} \to \mathbb{R} \) is the \textit{link function}. 
Single-index models

\[ \mathbf{X} \sim \text{Normal}(0, \mathbf{I}) \; ; \]
\[ Y \mid \mathbf{X} = \mathbf{x} \sim \text{Normal}(g(\langle \beta, \mathbf{x} \rangle), \sigma^2). \]

Here, \( g : \mathbb{R} \to \mathbb{R} \) is the link function.

- **Phase retrieval** (real signals): assume \( g(z) = z^2 \).
- **1-bit compressed sensing**: assume \( g(z) = \text{sign}(z) \).
- **Isotonic regression**: assume \( g \) is monotone (e.g., \( g' \geq 0 \)).
- **Convex regression**: assume \( g \) is convex (e.g., \( g'' \geq 0 \)).
- ...
Single-index models

\[
\begin{align*}
X & \sim \text{Normal}(0, I); \\
Y | X = x & \sim \text{Normal}(g(\langle \beta, x \rangle), \sigma^2).
\end{align*}
\]

Here, \( g: \mathbb{R} \to \mathbb{R} \) is the link function.

- **Phase retrieval** (real signals): assume \( g(z) = z^2 \).
- **1-bit compressed sensing**: assume \( g(z) = \text{sign}(z) \).
- **Isotonic regression**: assume \( g \) is monotone (e.g., \( g' \geq 0 \)).
- **Convex regression**: assume \( g \) is convex (e.g., \( g'' \geq 0 \)).
- \( \ldots \)

When \( g \) is unknown, model is generally called **single-index model**.
Single-index models

\[ X \sim \text{Normal}(0, I) ; \]
\[ Y \mid X = x \sim \text{Normal}(g(\langle \beta, x \rangle), \sigma^2) . \]

Here, \( g : \mathbb{R} \to \mathbb{R} \) is the \textit{link function}.

▶ \textbf{Phase retrieval} (real signals): assume \( g(z) = z^2 \).
▶ \textbf{1-bit compressed sensing}: assume \( g(z) = \text{sign}(z) \).
▶ \textbf{Isotonic regression}: assume \( g \) is monotone (e.g., \( g' \geq 0 \)).
▶ \textbf{Convex regression}: assume \( g \) is convex (e.g., \( g'' \geq 0 \)).
▶ ... 

When \( g \) is unknown, model is generally called \textbf{single-index model}.

\textbf{Semi-parametric estimation}: regard \( g \) as nuisance parameter; focus on estimating \( \beta \).
Aside: symmetric tensors and homogeneous polynomials

Recall formula for tensor function value:

\[ T(x^{(1)}, \ldots, x^{(p)}) = \sum_{i_1, \ldots, i_p} T_{i_1, \ldots, i_p} \cdot x_{i_1}^{(1)} \cdots x_{i_p}^{(p)}. \]
Aside: symmetric tensors and homogeneous polynomials

Recall formula for tensor function value:

\[ T(x^{(1)}, \ldots, x^{(p)}) = \sum_{i_1, \ldots, i_p} T_{i_1, \ldots, i_p} \cdot x^{(1)}_{i_1} \cdots x^{(p)}_{i_p}. \]

If \( T \) is symmetric (i.e., \( T_{i_1, \ldots, i_p} = T_{\pi(i_1), \ldots, \pi(i_p)} \) for any permutation \( \pi \)), then evaluating at \( x^{(1)} = \cdots = x^{(p)} = x \) gives

\[ T(x, \ldots, x) = p! \sum_{i_1 < \cdots < i_p} T_{i_1, \ldots, i_p} \cdot x_{i_1} \cdots x_{i_p}, \]

which is just the formula for a degree-\( p \) homogeneous polynomial.
Aside: symmetric tensors and homogeneous polynomials

Recall formula for tensor function value:

$$T(x^{(1)}, \ldots, x^{(p)}) = \sum_{i_1, \ldots, i_p} T_{i_1, \ldots, i_p} \cdot x_{i_1}^{(1)} \cdots x_{i_p}^{(p)}.$$  

If $T$ is symmetric (i.e., $T_{i_1, \ldots, i_p} = T_{\pi(i_1), \ldots, \pi(i_p)}$ for any permutation $\pi$), then evaluating at $x^{(1)} = \cdots = x^{(p)} = x$ gives

$$T(x, \ldots, x) = p! \sum_{i_1 < \cdots < i_p} T_{i_1, \ldots, i_p} \cdot x_{i_1} \cdots x_{i_p},$$

which is just the formula for a degree-$p$ homogeneous polynomial.

$p$-th order symmetric tensors $\simeq$ degree-$p$ homogeneous polynomials.
Let $H_p : \mathbb{R} \to \mathbb{R}$ denote the degree-$p$ Hermite polynomial.

Assume (for $Z \sim \text{Normal}(0, 1)$):

1. $\mathbb{E}[g(Z)^2] = 1$ (normalization—this is WLOG);
2. $\mathbb{E}[g'(Z)^2] \geq \epsilon$ (necessary for identifiability);
3. $g$ is smooth and $\mathbb{E}[g''(Z)^2] = O(1)$.
Using orthogonal polynomials  
(Dudeja & H., 2018)

Let $H_p : \mathbb{R} \to \mathbb{R}$ denote the degree-$p$ Hermite polynomial. Assume (for $Z \sim \text{Normal}(0, 1)$):

- $\mathbb{E}[g(Z)^2] = 1$ (normalization—this is WLOG);
- $\mathbb{E}[g'(Z)^2] \geq \epsilon$ (necessary for identifiability);
- $g$ is smooth and $\mathbb{E}[g''(Z)^2] = O(1)$.

There exists $p = O(1/\epsilon)$ such that

$$\mathbb{E}[Y H_p(\langle v, X \rangle)] = (\lambda \beta \otimes^p)(v), \quad v \in \mathbb{R}^d$$

for some $\lambda \neq 0$ with $|\lambda| = \Omega(\epsilon/\sqrt{p})$. 

Using orthogonal polynomials
(Dudeja & H., 2018)

Let $H_p : \mathbb{R} \rightarrow \mathbb{R}$ denote the degree-$p$ Hermite polynomial.

Assume (for $Z \sim \text{Normal}(0, 1)$):

- $\mathbb{E}[g(Z)^2] = 1$ (normalization—this is WLOG);
- $\mathbb{E}[g'(Z)^2] \geq \epsilon$ (necessary for identifiability);
- $g$ is smooth and $\mathbb{E}[g''(Z)^2] = O(1)$.

There exists $p = O(1/\epsilon)$ such that

$$\mathbb{E}[Y H_p(\langle \mathbf{v}, \mathbf{X} \rangle)] = (\lambda \beta^{\otimes p})(\mathbf{v}), \quad \mathbf{v} \in \mathbb{R}^d$$

for some $\lambda \neq 0$ with $|\lambda| = \Omega(\epsilon/\sqrt{p})$.

$\Rightarrow$ Get efficient algorithms for semi-parametric estimation of single-index model parameters, for very general link functions.
Recap

- Parameters of many latent variable models (satisfying non-degeneracy conditions) can be efficiently recovered from $O(1)$-order moments.
Recap

- Parameters of many latent variable models (satisfying non-degeneracy conditions) can be efficiently recovered from $O(1)$-order moments.
- Exploit distributional properties, multi-view structure, and other structure to determine usable moments.
Recap

- Parameters of many latent variable models (satisfying non-degeneracy conditions) can be efficiently recovered from $O(1)$-order moments.

- Exploit distributional properties, multi-view structure, and other structure to determine usable moments.

- Estimation via method-of-moments:
  1. Estimate moments $\rightarrow$ empirical moment tensor $\hat{T}$.
  2. Approximately decompose $\hat{T} \rightarrow$ parameter estimate $\hat{\theta}$. 
3. Error analysis
Moment estimates

Estimation of $\mathbb{E}[X^{\otimes 3}]$ (say) from iid sample $\{x_i\}_{i=1}^n$:

$$\hat{\mathbb{E}}[X^{\otimes 3}] := \frac{1}{n} \sum_{i=1}^{n} x_i^{\otimes 3}.$$
Moment estimates

Estimation of $\mathbb{E}[X^{\otimes 3}]$ (say) from iid sample $\{x_i\}_{i=1}^n$:

$$\widehat{\mathbb{E}[X^{\otimes 3}]} := \frac{1}{n} \sum_{i=1}^n x_i^{\otimes 3}.$$  

Inevitably expect error of order $n^{-1/2}$ in some norm, e.g.,

$$\|T\| := \sup_{\|x\|=\|y\|=\|z\|=1} T(x, y, z) \quad \text{(injective/"spectral" norm)},$$

$$\|T\|_F := \left( \sum_{i,j,k} T_{i,j,k}^2 \right)^{1/2} \quad \text{(Frobenius norm)}.$$
Nearly orthogonally decomposable tensor
(Mu, H., & Goldfarb, 2015)

Let $\varepsilon = \|E\|$ for $E := \hat{T} - T$.

Claim: Let $\hat{v} := \arg \max_{\|x\|=1} \hat{T}(x, x, x)$ and $\hat{\lambda} := \hat{T}(\hat{v}, \hat{v}, \hat{v})$.

Then

$$|\hat{\lambda} - \lambda_t| \leq \varepsilon, \quad \|\hat{v} - v_t\| \leq O \left( \frac{\varepsilon}{\lambda_t} + \left( \frac{\varepsilon}{\lambda_t} \right)^2 \right)$$

for some $t \in [d]$ with $\lambda_t \geq \max_{t'} \lambda_{t'} - 2\varepsilon$. 

Many efficient algorithms for solving this approximately, when $\varepsilon$ is small enough, like $1/d$ or $1/\sqrt{d}$ (e.g., Anandkumar, Ge, H., Kakade, & Telgarsky, 2014; Ma, Shi, & Steurer, 2016).
Nearly orthogonally decomposable tensor
(Mu, H., & Goldfarb, 2015)

Let \( \varepsilon = \|E\| \) for \( E := \hat{T} - T \).

**Claim:** Let \( \hat{v} := \arg \max_{\|x\|=1} \hat{T}(x, x, x) \) and \( \hat{\lambda} := \hat{T}(\hat{v}, \hat{v}, \hat{v}) \). Then

\[
|\hat{\lambda} - \lambda_t| \leq \varepsilon, \quad \|\hat{v} - v_t\| \leq O\left(\frac{\varepsilon}{\lambda_t} + \left(\frac{\varepsilon}{\lambda_t}\right)^2\right)
\]

for some \( t \in [d] \) with \( \lambda_t \geq \max_{t'} \lambda_{t'} - 2\varepsilon \).

Many efficient algorithms for solving this approximately, when \( \varepsilon \) is small enough, like \( 1/d \) or \( 1/\sqrt{d} \) (e.g., Anandkumar, Ge, H., Kakade, & Telgarsky, 2014; Ma, Shi, & Steurer, 2016).
Recall: greedy decomposition

(Zhang & Golub, 2001)

Matching moments:

\[
\left\{ (\hat{v}_t, \hat{\lambda}_t) \right\}_{t=1}^{d} := \arg \min_{\left\{ (x_t, \sigma_t) \right\}_{t=1}^{d}} \left\| T - \sum_{t=1}^{d} \sigma_t \cdot x_t \otimes x_t \otimes x_t \right\|_F^2.
\]
Recall: greedy decomposition
(Zhang & Golub, 2001)

Matching moments:

\[
\{(\hat{v}_t, \hat{\lambda}_t)\}_{t=1}^d := \arg\min\left\{ \left( x_t, \sigma_t \right) \right\} \left| \left| T - \sum_{t=1}^d \sigma_t \cdot x_t \otimes x_t \otimes x_t \right| \right|_F^2 .
\]

▶ Greedy approach:

▶ Find best rank-1 approximation:

\[
(\hat{v}, \hat{\lambda}) := \arg\min_{\|x\|=1, \sigma \geq 0} \left| \left| T - \sigma \cdot x \otimes x \otimes x \right| \right|_F^2 .
\]

▶ “Deflate”  

\[
T := T - \hat{\lambda} \cdot \hat{v} \otimes \hat{v} \otimes \hat{v}
\]

and repeat.
Recall: greedy decomposition

(Zhang & Golub, 2001)

**Matching moments:**

\[
\{(\hat{v}_t, \hat{\lambda}_t)\}_{t=1}^d := \arg\min_{\{(x_t, \sigma_t)\}_{t=1}^d} \left\| T - \sum_{t=1}^d \sigma_t \cdot x_t \otimes x_t \otimes x_t \right\|_F^2.
\]

▶ Greedy approach:

▶ Find best rank-1 approximation:

\[
\hat{v} := \arg\max_{\|x\|=1} T(x, x, x), \quad \hat{\lambda} := T(\hat{v}, \hat{v}, \hat{v}).
\]

▶ “Deflate”\( T := T - \hat{\lambda} \cdot \hat{v} \otimes \hat{v} \otimes \hat{v} \) and repeat.
Errors from deflation

(For simplicity, assume $\lambda_t = 1$ for all $t$, so $T = \sum_t v_t^{\otimes 3}$.)

**First greedy step:**
Rank-1 approx. $\hat{v}_1^{\otimes 3}$ to $\hat{T}$ satisfies $\|\hat{v}_1 - v_1\| \leq \varepsilon$ (say).
Errors from deflation

(For simplicity, assume $\lambda_t = 1$ for all $t$, so $T = \sum_t v_t \otimes^3$.)

**First greedy step:**
Rank-1 approx. $\hat{v}_1 \otimes^3$ to $\hat{T}$ satisfies $\|\hat{v}_1 - v_1\| \leq \varepsilon$ (say).

**Deflation:** To find next $v_t$, use

$$\hat{T} - \hat{v}_1 \otimes^3 = T + E - \hat{v}_1 \otimes^3$$

$$= \sum_{t=2}^{d} v_t \otimes^3 + E + (v_1 \otimes^3 - \hat{v}_1 \otimes^3).$$
Errors from deflation

(For simplicity, assume $\lambda_t = 1$ for all $t$, so $T = \sum_t v_t^{\otimes 3}$.)

**First greedy step:**
Rank-1 approx. $\hat{v}_1^{\otimes 3}$ to $\hat{T}$ satisfies $\|\hat{v}_1 - v_1\| \leq \varepsilon$ (say).

**Deflation:** To find next $v_t$, use

$$\hat{T} - \hat{v}_1^{\otimes 3} = T + E - \hat{v}_1^{\otimes 3}$$

$$= \sum_{t=2}^d v_t^{\otimes 3} + E + (v_1^{\otimes 3} - \hat{v}_1^{\otimes 3}).$$

Now error seems to have doubled (i.e., of size $2\varepsilon$) ...
Effect of deflation errors

For any unit vector $x$ orthogonal to $v_1$:

$$\left\| \nabla_x \left\{ \left( v_1 \otimes^3 - \hat{v}_1 \otimes^3 \right)(x, x, x) \right\} \right\| = \left\| \langle v_1, x \rangle^2 v_1 - \langle \hat{v}_1, x \rangle^2 \hat{v}_1 \right\|$$
Effect of deflation errors

For any unit vector \( x \) orthogonal to \( v_1 \):

\[
\left\| \frac{1}{3} \nabla_x \left\{ \left( v_1 \otimes^3 - \hat{v}_1 \otimes^3 \right) (x, x, x) \right\} \right\| = \left\| \langle v_1, x \rangle^2 v_1 - \langle \hat{v}_1, x \rangle^2 \hat{v}_1 \right\|
\]
\[= \langle \hat{v}_1, x \rangle^2
\]

So effect of errors (original and from deflation) \( E + (v_1 \otimes^3 - \hat{v}_1 \otimes^3) \) in directions orthogonal to \( v_1 \) is \((1 + o(1)) \varepsilon \) rather than \( 2 \varepsilon \).

Deflation errors have lower-order effect on finding other \( v_t \).

(Analogous statement for deflation with matrices does not hold.)
Effect of deflation errors

For any unit vector \( \mathbf{x} \) orthogonal to \( \mathbf{v}_1 \):

\[
\left\| \frac{1}{3} \nabla \left\{ \left( \mathbf{v}_1 \otimes^3 - \hat{\mathbf{v}}_1 \otimes^3 \right) \right\} \left( \mathbf{x}, \mathbf{x}, \mathbf{x} \right) \right\| = \left\| \left\langle \mathbf{v}_1, \mathbf{x} \right\rangle^2 \mathbf{v}_1 - \left\langle \hat{\mathbf{v}}_1, \mathbf{x} \right\rangle^2 \hat{\mathbf{v}}_1 \right\|
\]

\[
= \left\langle \hat{\mathbf{v}}_1, \mathbf{x} \right\rangle^2
\]

\[
\leq \left\| \mathbf{v}_1 - \hat{\mathbf{v}}_1 \right\|^2 \leq \varepsilon^2.
\]
Effect of deflation errors

For any unit vector $x$ orthogonal to $v_1$:

$$\left\| \frac{1}{3} \nabla x \left\{ \left( v_1^3 - \hat{v}_1^3 \right) (x, x, x) \right\} \right\| = \left\| \left< v_1, x \right>^2 v_1 - \left< \hat{v}_1, x \right>^2 \hat{v}_1 \right\|$$

$$= \left< \hat{v}_1, x \right>^2$$

$$\leq \left\| v_1 - \hat{v}_1 \right\|^2 \leq \varepsilon^2 .$$

So effect of errors (original and from deflation) $E + \left( v_1^3 - \hat{v}_1^3 \right)$ in directions orthogonal to $v_1$ is $(1 + o(1))\varepsilon$ rather than $2\varepsilon$. 

Deflation errors have lower-order effect on finding other $v_t$.

(Analogous statement for deflation with matrices does not hold.)
Effect of deflation errors

For any unit vector $\mathbf{x}$ orthogonal to $\mathbf{v}_1$:

\[
\left\| \frac{1}{3} \nabla \mathbf{x} \left\{ \left( \mathbf{v}_1^3 - \hat{\mathbf{v}}_1^3 \right) (\mathbf{x}, \mathbf{x}, \mathbf{x}) \right\} \right\| = \left\| \langle \mathbf{v}_1, \mathbf{x} \rangle^2 \mathbf{v}_1 - \langle \hat{\mathbf{v}}_1, \mathbf{x} \rangle^2 \hat{\mathbf{v}}_1 \right\|
\]

\[
= \langle \hat{\mathbf{v}}_1, \mathbf{x} \rangle^2
\]

\[
\leq \| \mathbf{v}_1 - \hat{\mathbf{v}}_1 \|^2 \leq \varepsilon^2.
\]

So effect of errors (original and from deflation) $E + (\mathbf{v}_1^3 - \hat{\mathbf{v}}_1^3)$ in directions orthogonal to $\mathbf{v}_1$ is $(1 + o(1))\varepsilon$ rather than $2\varepsilon$.

- Deflation errors have lower-order effect on finding other $\mathbf{v}_t$.

  (Analogous statement for deflation with matrices does not hold.)
Summary

- Using method-of-moments with low-order moments, can efficiently estimate parameters for many models.
  - Exploit distributional properties, multi-view structure, and other structure to determine usable moments tensors.
  - Some efficient algorithms for carrying out the tensor decomposition to obtain parameter estimates.
Summary

- Using method-of-moments with low-order moments, can efficiently estimate parameters for many models.
  - Exploit distributional properties, multi-view structure, and other structure to determine usable moments tensors.
  - Some efficient algorithms for carrying out the tensor decomposition to obtain parameter estimates.

- Many issues to resolve!
  - Handle model misspecification, increase robustness.
  - General methodology.
  - Incorporate general prior knowledge and interactive feedback.
Acknowledgements

Collaborators: Anima Anandkumar (Caltech), Rishabh Dudeja (Columbia), Dean Foster (Amazon), Rong Ge (Duke), Don Goldfarb (Columbia), Sham Kakade (UW), Percy Liang (Stanford), Yi-Kai Liu (NIST), Cun Mu (Jet), Matus Telgarsky (UIUC), Tong Zhang (Tencent)

Further reading:

▶ Anandkumar, Ge, H., Kakade, & Telgarsky. 
Tensor decompositions for learning latent variable models. 
https://goo.gl/F8HudN

¡Gracias!