# Learning Overcomplete Latent Variable Models through Tensor Decompositions

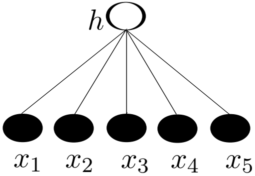
**Anima Anandkumar**

U.C. Irvine

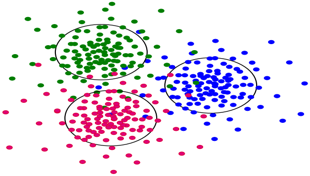
Joint work with Rong Ge and Majid Janzamin.

# Tensor-Based Learning

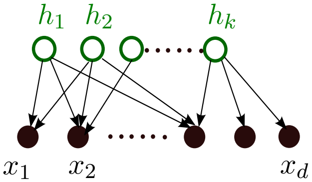
## Multi-view mixtures




## Spherical Gaussian mixtures

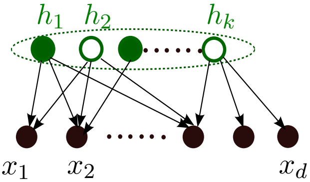


## Indep. Component Analysis




Independent $h_i$'s

## Sparse Coding




Sparse $h_i$'s

# General Framework

- Discover hidden structure in data: unsupervised and semi-supervised learning of latent variable models.
- Moment-based estimation: Compute low order moments (up to fourth order) from observed data.

## In this talk

- Unsupervised and semi-supervised learning through tensor decomposition
- Overcomplete models: Number of latent components greater than observed dimension.
- Tight sample complexity bounds: Novel concentration bounds for tensors.

# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.

# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.
- For tensor $T$, find decomposition into rank one terms

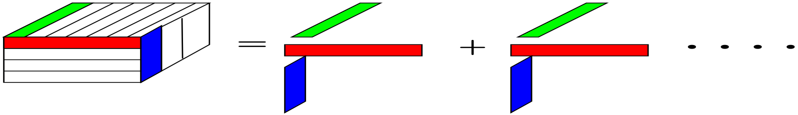
$$\boxed{T = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j, \quad a_j, b_j, c_j \in \mathcal{S}^{d-1}.}$$

# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.
- For tensor $T$, find decomposition into rank one terms

$$\boxed{T = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j, \quad a_j, b_j, c_j \in \mathcal{S}^{d-1}.}$$



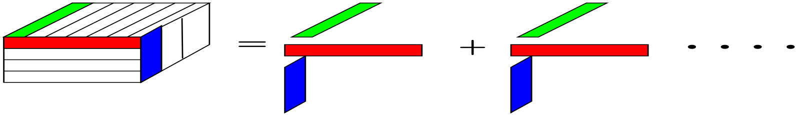
Tensor $T$ $\quad$ $w_1 \cdot a_1 \otimes b_1 \otimes c_1$ $\quad$ $w_2 \cdot a_2 \otimes b_2 \otimes c_2$

# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.
- For tensor $T$, find decomposition into rank one terms

$$\boxed{T = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j, \quad a_j, b_j, c_j \in \mathcal{S}^{d-1}.}$$



Tensor $T$ $\qquad$ $w_1 \cdot a_1 \otimes b_1 \otimes c_1$ $\qquad$ $w_2 \cdot a_2 \otimes b_2 \otimes c_2$

- $k$: tensor rank, $d$: ambient dimension.
- $k \leq d$: undercomplete and $k > d$: overcomplete.

In this talk: guarantees for overcomplete tensor decomposition

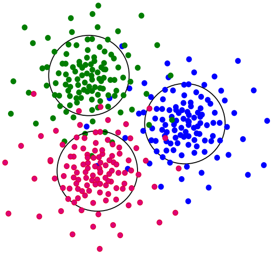
# Outline


1. Introduction
2. Summary of Results
3. Tensor Decomposition
4. Guarantees for Alternating Minimization
5. Conclusion and Other Results

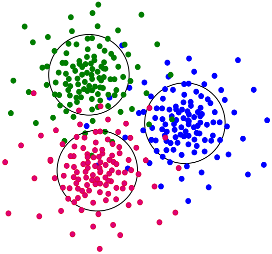
# Spherical Gaussian Mixtures

## Assumptions

- $k$ components, $d$: observed dimension.
- Component means $a_i$ incoherent: randomly drawn from the sphere.
- Spherical variance $\frac{\sigma^2}{d} I$ (assume known).



# Spherical Gaussian Mixtures



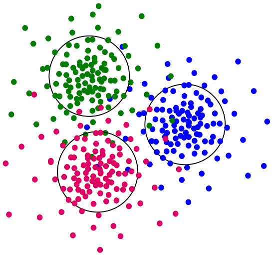
## Assumptions

- $k$ components, $d$: observed dimension.
- Component means $a_i$ incoherent: randomly drawn from the sphere.
- Spherical variance $\frac{\sigma^2}{d} I$ (assume known).

## In this talk: special case

- Noise norm $\sigma^2 = 1$: same as signal.
- Uniform probability of components.

# Spherical Gaussian Mixtures



## Assumptions

- $k$ components, $d$: observed dimension.
- Component means $a_i$ incoherent: randomly drawn from the sphere.
- Spherical variance $\frac{\sigma^2}{d} I$ (assume known).

## In this talk: special case

- Noise norm $\sigma^2 = 1$: same as signal.
- Uniform probability of components.

## Tensor For Learning (Hsu, Kakade 2012)

$$M_3 := \mathbb{E}[x^{\otimes 3}] - \sigma^2 \sum_{i \in [d]} (\mathbb{E}[x] \otimes e_i \otimes e_i + \dots)$$

# Semi-supervised Learning of Gaussian Mixtures

- $n$ unlabeled samples, $m_j$: samples for component $j$.
- No. of mixture components: $k = o(d^{1.5})$
- No. of labeled samples: $m_j = \tilde{\Omega}(1)$.
- No. of unlabeled samples: $n = \tilde{\Omega}(k)$.

Our result: achieved error with $n$ unlabeled samples

$$\boxed{\max_i \|\widehat{a}_i - a_i\| = \tilde{O}\left(\sqrt{\frac{k}{n}}\right) + \tilde{O}\left(\frac{\sqrt{k}}{d}\right)}$$

- Can handle (polynomially) overcomplete mixtures.
- Extremely small number of labeled samples: $\mathrm{polylog}(d)$.
- Sample complexity is tight: need $\tilde{\Omega}(k)$ samples!
- Approximation error: decaying in high dimensions.

# Unsupervised Learning of Gaussian Mixtures

## Conditions for recovery

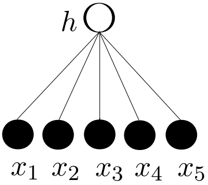
- No. of mixture components: $k = C \cdot d$
- No. of unlabeled samples: $n = \tilde{\Omega}(k \cdot d)$.
- Computational complexity: $\tilde{O}\left(e^{C^2}\right)$

## Our result: achieved error with $n$ unlabeled samples

$$\boxed{\max_i \|\widehat{a}_i - a_i\| = \tilde{O}\left(\sqrt{\frac{k}{n}}\right) + \tilde{O}\left(\frac{\sqrt{k}}{d}\right)}$$

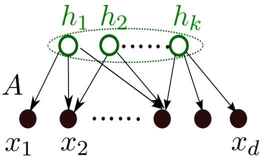
- Error: same as before, for semi-supervised setting.
- Sample complexity: worse than semi-supervised, but better than previous works (no dependence on condition number of $A$).
- Computational complexity: polynomial when $k = \Theta(d)$.

# Multi-view Mixture Models




- Linear model: $x_i = A_i h + z_i$.
- Incoherence: The columns of $A_i$ are incoherent (randomly drawn from sphere).
- The noise $z_i$ satisfy RIP, e.g. Gaussian, Bernoulli.
- Same results as Gaussian mixtures.

# Independent Component Analysis



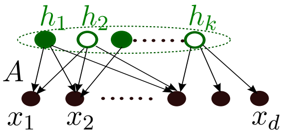

- Independent sources, unknown mixing.
- Blind source separation of speech, image, video..
- Form cumulant tensor $M_4 := \mathbb{E}[x^{\otimes 4}] - \dots$
- $n$ samples. $k$ sources. $d$ dimensions.
- $x = Ah$. Columns of $A$ are incoherent.
- Sources $h$ are kurtotic.

## Learning Result

- Semi-supervised: $k = o(d^2)$, $n \geq \max(k^2, k^4/d^3)$.
- Unsupervised: $k = O(d)$, $n \geq k^3$.

$$\boxed{\max_i \min_{f \in \{-1,1\}} \|f\widehat{a}_i - a_i\| = \tilde{O}\left(\sqrt{\frac{k^2}{\min\left(n, \sqrt{d^3 n}\right)}}\right) + \tilde{O}\left(\frac{\sqrt{k}}{d^{1.5}}\right)}$$

# Sparse Coding




- Sparse coefficients, unknown dictionary.
- Image compression, feature learning...
- $x = Ah$. Columns of $A$ are incoherent.
  - Coefficients $h$ are independent Bernoulli Gaussian: Sparse ICA.
  - Form cumulant tensor $M_4 := \mathbb{E}[x^{\otimes 4}] - \ldots$
  - $n$ samples. $k$ dictionary elements. $d$ dimensions. $s$ avg. sparsity.

## Learning Result

- Semi-supervised: $k = o(d^2)$, $n \geq \max(sk, s^2k^2/d^3)$.
- Unsupervised: $k = O(d)$, $n \geq sk^2$.

$$\boxed{\max_i \min_{f \in \{-1,1\}} \|f\widehat{a}_i - a_i\| = \tilde{O}\left(\sqrt{\frac{sk}{\min\left(n, \sqrt{d^3 n}\right)}}\right) + \tilde{O}\left(\frac{\sqrt{k}}{d^{1.5}}\right)}$$

# Outline

1 Introduction

2 Summary of Results

3 Tensor Decomposition

4 Guarantees for Alternating Minimization

5 Conclusion and Other Results

# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.

# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.
- For tensor $T$, find decomposition into rank one terms

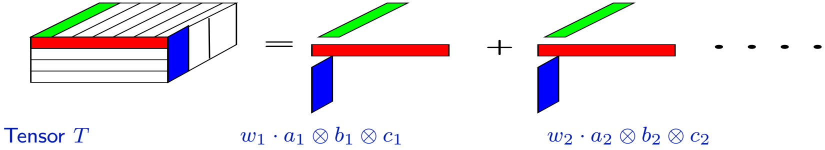
$$\boxed{T = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j, \quad a_j, b_j, c_j \in \mathcal{S}^{d-1}.}$$

# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.
- For tensor $T$, find decomposition into rank one terms

$$\boxed{T = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j, \quad a_j, b_j, c_j \in \mathcal{S}^{d-1}.}$$

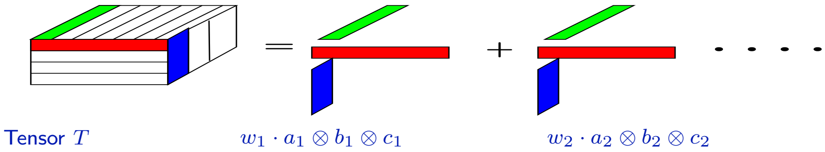


# Tensor Decomposition

## CANDECOMP/PARAFAC (CP) Decomposition

- $a \otimes b \otimes c$ is a rank-1 tensor whose $\mathbf{i}^{\text{th}}$ entry is $a(i_1) \cdot b(i_2) \cdot c(i_3)$.
- For tensor $T$, find decomposition into rank one terms

$$\boxed{T = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j, \quad a_j, b_j, c_j \in \mathcal{S}^{d-1}.}$$




- $k$: tensor rank, $d$: ambient dimension.
- $k \leq d$: undercomplete and $k > d$: overcomplete.

In this talk: guarantees for overcomplete tensor decomposition

# Background on Tensor Decomposition

$$\boxed{T = \sum_{i \in [k]} w_i a_i \otimes b_i \otimes c_i, \quad a_i, b_i, c_i \in \mathcal{S}^{d-1}.}$$

## Theoretical Guarantees

- Tensor decompositions in psychometrics (Cattell '44).
- CP tensor decomposition (Harshman '70, Carol & Chang '70).
- Identifiability of CP tensor decomposition (Kruskal '76).
- Orthogonal decomposition: (Zhang & Golub '01, Kolda '01).
- Tensor decomposition through (lifted) linear equations (Lawthauwer '07): works for overcomplete tensors.
- Tensor decomposition through simultaneous diagonalization: perturbation analysis (Goyal et. al '13, Bhaskara '13)

# Background on Tensor Decompositions (contd.)

$$\boxed{T = \sum_{i \in [k]} w_i a_i \otimes b_i \otimes c_i, \quad a_i, b_i, c_i \in \mathcal{S}^{d-1}.}$$

Practice: Alternating least squares (ALS)

- Let $A = [a_1 | a_2 \ldots a_k]$ and similarly $B, C$.
- Fix estimates of two of the modes (say for $A$ and $B$) and re-estimate the third.
- Iterative updates, low computational complexity.
- No theoretical guarantees.

In this talk: analysis of alternating minimization

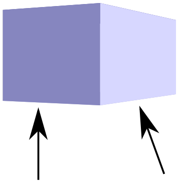
# Alternating Minimization

$$\boxed{T = \sum_{i \in [k]} w_i a_i \otimes b_i \otimes c_i, \quad a_i, b_i, c_i \in \mathcal{S}^{d-1}.}$$

## Rank-1 Updates

- Initialization: $a^{(0)}, b^{(0)}, c^{(0)}$.
- Update in $t^{\text{th}}$ step: fix $a^{(t)}, b^{(t)}$ and

$$c^{(t)} \propto T(a^{(t)}, b^{(t)}, I) = \sum_{i \in [k]} w_i \langle a_i, a^{(t)} \rangle \langle b_i, b^{(t)} \rangle c_i.$$



- After (approx.) convergence, restart.

# Optimization Viewpoint

## Best Rank-1 Approximation

$$\boxed{\min_{a,b,c \in \mathcal{S}^{d-1}, w \in \mathbb{R}} \|T - w \cdot a \otimes b \otimes c\|_F.}$$

## Challenges

- Optimization problem: non-convex, multiple local optima.

# Optimization Viewpoint

## Best Rank-1 Approximation

$$\boxed{\min_{a,b,c\in\mathcal{S}^{d-1},w\in\mathbb{R}} \|T - w \cdot a \otimes b \otimes c\|_F.}$$

## Challenges

- Optimization problem: non-convex, multiple local optima.
- Alternating minimization: improves above objective in each step.

# Optimization Viewpoint

## Best Rank-1 Approximation

$$\boxed{\min_{a,b,c\in\mathcal{S}^{d-1},w\in\mathbb{R}} \|T - w\cdot a\otimes b\otimes c\|_F.}$$

## Challenges

- Optimization problem: non-convex, multiple local optima.
- Alternating minimization: improves above objective in each step.
- Recovery of $a_i, b_i, c_i$'s?

# Optimization Viewpoint

## Best Rank-1 Approximation

$$\boxed{\min_{a,b,c\in\mathcal{S}^{d-1},w\in\mathbb{R}} \|T - w\cdot a\otimes b\otimes c\|_F.}$$

## Challenges

- Optimization problem: non-convex, multiple local optima.
- Alternating minimization: improves above objective in each step.
- Recovery of $a_i, b_i, c_i$'s? Not true in general.

# Optimization Viewpoint

## Best Rank-1 Approximation

$$\boxed{\min_{a,b,c \in \mathcal{S}^{d-1}, w \in \mathbb{R}} \|T - w \cdot a \otimes b \otimes c\|_F.}$$

## Challenges

- Optimization problem: non-convex, multiple local optima.
- Alternating minimization: improves above objective in each step.
- Recovery of $a_i, b_i, c_i$'s? Not true in general.
- Noisy tensor decomposition: exact $T$ not available, robustness? sample complexity?

Natural conditions under which Alt-Min has guarantees?

# Special case: Orthogonal Setting

- $\langle a_i, a_j \rangle = 0$, for $i \neq j$. Similarly for $b, c$.
- Alternating updates:

$$c^{(t)} \propto T(a^{(t)}, b^{(t)}, I) = \sum_{i \in [k]} w_i \langle a_i, a^{(t)} \rangle \langle b_i, b^{(t)} \rangle c_i.$$

- $a_i, b_i, c_i$ are stationary points.

---

"Tensor Decompositions for Learning Latent Variable Models" by A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade and M. Telgarsky. Preprint, October 2012.

# Special case: Orthogonal Setting

- $\langle a_i, a_j \rangle = 0$, for $i \neq j$. Similarly for $b, c$.
- Alternating updates:

$$c^{(t)} \propto T(a^{(t)}, b^{(t)}, I) = \sum_{i \in [k]} w_i \langle a_i, a^{(t)} \rangle \langle b_i, b^{(t)} \rangle c_i.$$

- $a_i, b_i, c_i$ are stationary points.
- ONLY local optima for best rank-1 approximation problem.
- Guaranteed recovery through alternating minimization.

---

"Tensor Decompositions for Learning Latent Variable Models" by A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade and M. Telgarsky. Preprint, October 2012.

# Special case: Orthogonal Setting

- $\langle a_i, a_j \rangle = 0$, for $i \neq j$. Similarly for $b, c$.
- Alternating updates:

$$c^{(t)} \propto T(a^{(t)}, b^{(t)}, I) = \sum_{i \in [k]} w_i \langle a_i, a^{(t)} \rangle \langle b_i, b^{(t)} \rangle c_i.$$

- $a_i, b_i, c_i$ are stationary points.
- ONLY local optima for best rank-1 approximation problem.
- Guaranteed recovery through alternating minimization.
- Perturbation Analysis: Under poly$(d)$ number of random initializations and bounded noise conditions.

---

"Tensor Decompositions for Learning Latent Variable Models" by A. Anandkumar, R. Ge, D. Hsu, S.M. Kakade and M. Telgarsky. Preprint, October 2012.

# Beyond Orthogonal Tensor Decomposition

## Limitations

- Not ALL tensors have orthogonal decomposition (unlike matrices).

# Beyond Orthogonal Tensor Decomposition

## Limitations

- Not ALL tensors have orthogonal decomposition (unlike matrices).
- Orthogonal forms: cannot handle overcomplete tensors $(k > d)$.

# Beyond Orthogonal Tensor Decomposition

## Limitations

- Not ALL tensors have orthogonal decomposition (unlike matrices).
- Orthogonal forms: cannot handle overcomplete tensors $(k > d)$.
- Overcomplete representations: redundancy leads to flexible modeling, noise resistant, no domain knowledge.

# Beyond Orthogonal Tensor Decomposition

## Limitations

- Not ALL tensors have orthogonal decomposition (unlike matrices).
- Orthogonal forms: cannot handle overcomplete tensors $(k > d)$.
- Overcomplete representations: redundancy leads to flexible modeling, noise resistant, no domain knowledge.

## Undercomplete tensors $(k \leq d)$ with full rank components

- Assume $A, B, C$ have full column rank.
- Whitening: Compute multilinear transformation to obtain an orthogonal form.
- Limitations: depends on condition number, sensitive to noise.

# Our Setup

## So far

- General tensor decomposition: NP-hard.
- Orthogonal tensors: too limiting.

---

"Guaranteed Tensor Decomposition via Alternating Minimization" by M. Janzamin, A. Anandkumar, and R. Ge. Preprint, Jan 2014.

# Our Setup

## So far

- General tensor decomposition: NP-hard.
- Orthogonal tensors: too limiting.
- Tractable cases? Covers overcomplete tensors?

---

"Guaranteed Tensor Decomposition via Alternating Minimization" by M. Janzamin, A. Anandkumar, and R. Ge. Preprint, Jan 2014.

# Our Setup

## So far

- General tensor decomposition: NP-hard.
- Orthogonal tensors: too limiting.
- Tractable cases? Covers overcomplete tensors?

## Our framework: Incoherent Components

- $|\langle a_i, a_j \rangle| = O\left(1/\sqrt{d}\right)$ for $i \neq j$. Similarly for $b, c$.
- Can handle overcomplete tensors. Satisfied by random (generic) vectors.

---

"Guaranteed Tensor Decomposition via Alternating Minimization" by M. Janzamin, A. Anandkumar, and R. Ge. Preprint, Jan 2014.

# Our Setup

## So far

- General tensor decomposition: NP-hard.
- Orthogonal tensors: too limiting.
- Tractable cases? Covers overcomplete tensors?

## Our framework: Incoherent Components

- $|\langle a_i, a_j \rangle| = O\left(1/\sqrt{d}\right)$ for $i \neq j$. Similarly for $b, c$.
- Can handle overcomplete tensors. Satisfied by random (generic) vectors.

Guaranteed recovery for alternating minimization?

---

"Guaranteed Tensor Decomposition via Alternating Minimization" by M. Janzamin, A. Anandkumar, and R. Ge. Preprint, Jan 2014.

# Analysis of One Step Update

## Basic Intuition

- Let $\hat{a}, \hat{b}$ be "close to" $a_1, b_1$. Alternating update:

$$\begin{aligned}\hat{c} \propto T(\hat{a}, \hat{b}, I) &= \sum_{i \in [k]} w_i \langle a_i, \hat{a} \rangle \langle b_i, \hat{b} \rangle c_i, \\ &= w_1 \langle a_1, \hat{a} \rangle \langle b_1, \hat{b} \rangle + T_{-1}(\hat{a}, \hat{b}, I).\end{aligned}$$

# Analysis of One Step Update

## Basic Intuition

- Let $\hat{a}, \hat{b}$ be "close to" $a_1, b_1$. Alternating update:

$$\hat{c} \propto T(\hat{a}, \hat{b}, I) = \sum_{i \in [k]} w_i \langle a_i, \hat{a} \rangle \langle b_i, \hat{b} \rangle c_i,$$
$$= w_1 \langle a_1, \hat{a} \rangle \langle b_1, \hat{b} \rangle + T_{-1}(\hat{a}, \hat{b}, I).$$

- $T_{-1}(\hat{a}, \hat{b}, I) = 0$ in orthogonal case, when $\hat{a} = a_1, \hat{b} = b_1$.

# Analysis of One Step Update

## Basic Intuition

- Let $\hat{a}, \hat{b}$ be "close to" $a_1, b_1$. Alternating update:

$$\begin{aligned}\hat{c} \propto T(\hat{a}, \hat{b}, I) &= \sum_{i \in [k]} w_i \langle a_i, \hat{a} \rangle \langle b_i, \hat{b} \rangle c_i, \\ &= w_1 \langle a_1, \hat{a} \rangle \langle b_1, \hat{b} \rangle + T_{-1}(\hat{a}, \hat{b}, I).\end{aligned}$$

- $T_{-1}(\hat{a}, \hat{b}, I) = 0$ in orthogonal case, when $\hat{a} = a_1, \hat{b} = b_1$.
- Can it be controlled for incoherent (random) vectors?

# Outline


1. Introduction
2. Summary of Results
3. Tensor Decomposition
4. Guarantees for Alternating Minimization
5. Conclusion and Other Results

# Results for one step update

- Incoherence: $|\langle a_i, a_j \rangle| = O\left(1/\sqrt{d}\right)$ for $i \neq j$. Similarly for $b, c$.
- Spectral norm: $\|A\|, \|B\|, \|C| \leq 1 + O\left(\sqrt{\frac{k}{d}}\right)$. $\|T\| \leq (1 + o(1))$.
- Tensor rank: $k = o(d^{1.5})$. Weights: For simplicity, $w_i \equiv 1$.
- $\mathrm{dist}(\hat{a}, a) := \min_f \|f\hat{a} - a\|$ for normalized $\hat{a}, a$.

# Results for one step update

- Incoherence: $|\langle a_i, a_j \rangle| = O\left(1/\sqrt{d}\right)$ for $i \neq j$. Similarly for $b, c$.
- Spectral norm: $\|A\|, \|B\|, \|C| \leq 1 + O\left(\sqrt{\frac{k}{d}}\right)$. $\|T\| \leq (1 + o(1))$.
- Tensor rank: $k = o(d^{1.5})$. Weights: For simplicity, $w_i \equiv 1$.
- $\mathrm{dist}(\hat{a}, a) := \min_f \|f\hat{a} - a\|$ for normalized $\hat{a}, a$.

Lemma (AGJ 2014)

$\mathrm{dist}(a_1, \hat{a}) \leq \epsilon$, similarly for $\hat{b}$, and $1 - \epsilon^2 > f(\epsilon; k, d)$, after one step

$$\boxed{\mathrm{dist}(\hat{c}, c_1) \leq \frac{f(\epsilon; k, d)}{1 - \epsilon^2 - f(\epsilon; k, d)}.}$$

# Results for one step update

- Incoherence: $|\langle a_i, a_j \rangle| = O\left(1/\sqrt{d}\right)$ for $i \neq j$. Similarly for $b, c$.
- Spectral norm: $\|A\|, \|B\|, \|C| \leq 1 + O\left(\sqrt{\frac{k}{d}}\right)$. $\|T\| \leq (1 + o(1))$.
- Tensor rank: $k = o(d^{1.5})$. Weights: For simplicity, $w_i \equiv 1$.
- $\mathrm{dist}(\hat{a}, a) := \min_f \|f\hat{a} - a\|$ for normalized $\hat{a}, a$.

Lemma (AGJ 2014)

$\mathrm{dist}(a_1, \hat{a}) \leq \epsilon$, similarly for $\hat{b}$, and $1 - \epsilon^2 > f(\epsilon; k, d)$, after one step

$$\boxed{\mathrm{dist}(\hat{c}, c_1) \leq \frac{f(\epsilon; k, d)}{1 - \epsilon^2 - f(\epsilon; k, d)}}.$$

- $f(\epsilon; k, d) := O\left(\frac{\sqrt{k}}{d} + \max\left(\frac{1}{\sqrt{d}}, \frac{k}{d^{1.5}}\right)\epsilon + \epsilon^2\right)$.
- $\frac{\sqrt{k}}{d}$: approximation error, rest: error contraction.

# Main Result: Local Convergence

- Initialization: $\mathrm{dist}(a_1, \hat{a}) \leq \epsilon_0$, similarly for $\hat{b}$ and $\epsilon_0 <$ const.
- Noise: $\hat{T} := T + E$, and $\|E\| \leq 1/\operatorname{polylog}(d)$.
- Approximation error: $\epsilon_T := \|E\| + \tilde{O}\left(\frac{\sqrt{k}}{d}\right)$

# Main Result: Local Convergence

- Initialization: $\mathrm{dist}(a_1, \hat{a}) \leq \epsilon_0$, similarly for $\hat{b}$ and $\epsilon_0 <$ const.
- Noise: $\hat{T} := T + E$, and $\|E\| \leq 1/\operatorname{polylog}(d)$.
- Approximation error: $\epsilon_T := \|E\| + \tilde{O}\left(\frac{\sqrt{k}}{d}\right)$

Theorem (Local Convergence)

After $O(\log(1/\epsilon_T))$ steps of alternating rank-1 updates,

$$\boxed{\mathrm{dist}(a_1, a^{(t)}) = O(\epsilon_T).}$$

# Main Result: Local Convergence

- Initialization: $\text{dist}(a_1, \hat{a}) \leq \epsilon_0$, similarly for $\hat{b}$ and $\epsilon_0 <$ const.
- Noise: $\hat{T} := T + E$, and $\|E\| \leq 1/\operatorname{polylog}(d)$.
- Approximation error: $\epsilon_T := \|E\| + \tilde{O}\left(\frac{\sqrt{k}}{d}\right)$

Theorem (Local Convergence)

After $O(\log(1/\epsilon_T))$ steps of alternating rank-1 updates,

$$\boxed{\text{dist}(a_1, a^{(t)}) = O(\epsilon_T).}$$

- Linear convergence: up to approximation error.

# Main Result: Local Convergence

- Initialization: $\text{dist}(a_1, \hat{a}) \leq \epsilon_0$, similarly for $\hat{b}$ and $\epsilon_0 < \text{const}$.
- Noise: $\hat{T} := T + E$, and $\|E\| \leq 1/\operatorname{polylog}(d)$.
- Approximation error: $\epsilon_T := \|E\| + \tilde{O}\left(\frac{\sqrt{k}}{d}\right)$

### Theorem (Local Convergence)

After $O(\log(1/\epsilon_T))$ steps of alternating rank-1 updates,

$$\boxed{\text{dist}(a_1, a^{(t)}) = O(\epsilon_T).}$$

- Linear convergence: up to approximation error.
- Guarantees for overcomplete tensors: $k = o(d^{1.5})$ and for $p^{\text{th}}$-order tensors $k = o(d^{p/2})$.

# Main Result: Local Convergence

- Initialization: $\mathrm{dist}(a_1, \hat{a}) \leq \epsilon_0$, similarly for $\hat{b}$ and $\epsilon_0 <$ const.
- Noise: $\hat{T} := T + E$, and $\|E\| \leq 1/\operatorname{polylog}(d)$.
- Approximation error: $\epsilon_T := \|E\| + \tilde{O}\left(\frac{\sqrt{k}}{d}\right)$

Theorem (Local Convergence)

After $O(\log(1/\epsilon_T))$ steps of alternating rank-1 updates,

$$\boxed{\mathrm{dist}(a_1, a^{(t)}) = O(\epsilon_T).}$$

- Linear convergence: up to approximation error.
- Guarantees for overcomplete tensors: $k = o(d^{1.5})$ and for $p^{\text{th}}$-order tensors $k = o(d^{p/2})$.
- Requires good initialization. What about global convergence?

# Global Convergence $k = O(d)$

## SVD Initialization

- Find the top singular vectors of $T(I, I, \theta)$ for $\theta \sim \mathcal{N}(0, I)$.
- Use them for initialization. $L$ trials.

# Global Convergence $k = O(d)$

## SVD Initialization

- Find the top singular vectors of $T(I, I, \theta)$ for $\theta \sim \mathcal{N}(0, I)$.
- Use them for initialization. $L$ trials.

## Assumptions

# Global Convergence $k = O(d)$

## SVD Initialization

- Find the top singular vectors of $T(I, I, \theta)$ for $\theta \sim \mathcal{N}(0, I)$.
- Use them for initialization. $L$ trials.

## Assumptions

- Number of initializations: $L \geq k^{\Omega(k/d)^2}$, Tensor Rank: $k = O(d)$

# Global Convergence $k = O(d)$

## SVD Initialization

- Find the top singular vectors of $T(I, I, \theta)$ for $\theta \sim \mathcal{N}(0, I)$.
- Use them for initialization. $L$ trials.

## Assumptions

- Number of initializations: $L \geq k^{\Omega(k/d)^2}$, Tensor Rank: $k = O(d)$
- No. of Iterations: $N = \Theta\left(\log(1/\epsilon_T)\right)$. Recall $\epsilon_T$: approx. error.

# Global Convergence $k = O(d)$

## SVD Initialization

- Find the top singular vectors of $T(I, I, \theta)$ for $\theta \sim \mathcal{N}(0, I)$.
- Use them for initialization. $L$ trials.

## Assumptions

- Number of initializations: $L \geq k^{\Omega(k/d)^2}$, Tensor Rank: $k = O(d)$
- No. of Iterations: $N = \Theta\left(\log(1/\epsilon_T)\right)$. Recall $\epsilon_T$: approx. error.

Theorem (Global Convergence) $\boxed{\mathrm{dist}(a_1, a^{(N)}) \leq O(\epsilon_T).}$

# Global Convergence $k = O(d)$

## SVD Initialization

- Find the top singular vectors of $T(I, I, \theta)$ for $\theta \sim \mathcal{N}(0, I)$.
- Use them for initialization. $L$ trials.

## Assumptions

- Number of initializations: $L \geq k^{\Omega(k/d)^2}$, Tensor Rank: $k = O(d)$
- No. of Iterations: $N = \Theta\left(\log(1/\epsilon_T)\right)$. Recall $\epsilon_T$: approx. error.

Theorem (Global Convergence) $\boxed{\mathrm{dist}(a_1, a^{(N)}) \leq O(\epsilon_T).}$

## Corollary: Differing Dimensions

- If $a_i, b_i \in \mathbb{R}^{d_u}$ and $c_i \in \mathbb{R}^{d_o}$, and $d_u \geq k \geq d_o$.
- $k = O(\sqrt{d_u d_o})$ for incoherent vectors. $k = O(d_u)$ if $A, B$ orthogonal.
- Same guarantees. Can handle one overcomplete mode.

# High-level Intuition for Sample Bounds

- Multi-view Model: $x_1 = Ah + z_i$, where $z_i$ is noise.
- Exact moment $T = \sum_i w_i a_i \otimes b_i \otimes c_i$.
- Sample moment: $\hat{T} = \frac{1}{n} \sum_i x_1^i \otimes x_2^i \otimes x_3^i$.

Naive Idea: $\|\hat{T} - T\| \leq \|\operatorname{mat}(\hat{T}) - \operatorname{mat}(T)\|$, apply matrix Bernstein's.

- Our idea: Careful $\epsilon$-net covering for $\hat{T} - T$.
- $\hat{T} - T$ has many terms, e.g. all-noise term: $\frac{1}{n} \sum_i z_1^i \otimes z_2^i \otimes z_3^i$ and signal-noise terms.
- Need to bound $\displaystyle\frac{1}{n} \sum_i \langle z_1^i, u\rangle \langle z_2^i, v\rangle \langle z_3^i, w\rangle$, for all $u, v, w \in \mathcal{S}^{d-1}$.
- Classify inner products into buckets and bound them separately.

Tight sample bounds for a range of latent variable models

# Outline

1 Introduction

2 Summary of Results

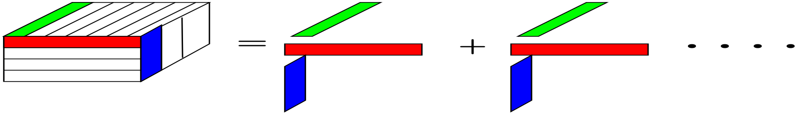
3 Tensor Decomposition

4 Guarantees for Alternating Minimization

5 Conclusion and Other Results

# Conclusion

$$\boxed{T = \sum_{i \in [k]} w_i a_i \otimes b_i \otimes c_i, \quad a_i, b_i, c_i \in \mathcal{S}^{d-1}.}$$



Tensor $T$ $\qquad$ $w_1 \cdot a_1 \otimes b_1 \otimes c_1$ $\qquad$ $w_2 \cdot a_2 \otimes b_2 \otimes c_2$

## Summary

- Analysis of alternating rank-1 updates under incoherent components.
- (Approx.) local convg. $k = o(d^{1.5})$, global convg. $k = O(d)$.
- Efficient learning and tight sample complexity for various latent variable models.

# Other Works on Tensor Decompositions

## Large-Scale Cloud Implementation on REEF

- F. Huang, N. Karampatziakis, S. Matusevych, P. Mineiro, A. Anandkumar, "Tensor Decompositions on REEF," under preparation.
- Code will soon be available.

## Parallelized Hierarchical Tensor Decomposition

- F. Huang, U. N. Niranjan, A. Anandkumar, "Integrated Structure and Parameter Learning in Latent Tree Graphical Models," on ArXiv.
- Code available at

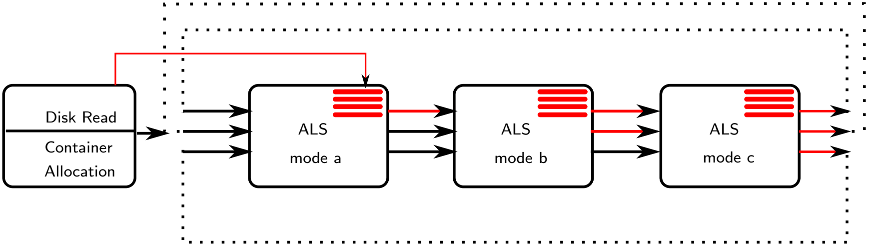
  `https://github.com/FurongHuang/StructureParameterLatentTree.git`
- Talk tomorrow at Learning Tractable Probabilistic Models (LTPM) workshop at 14:00.

# Tensor Factorization on REEF

## Large-scale implementation

- Map-Reduce: huge overhead in disk reading, container allocation.
- REEF: Retainable Evaluator Execution Framework.
- Advantage: Open source distributed system with one time container allocation, keep the tensor in memory

## Solution: REEF



# Preliminary Evaluation

## New York Times Corpus

- Documents $n = 300,000$
- Vocabulary $d = 100,000$
- Topics $k = 100$

| | Stochastic Variational Inference | Tensor Decomposition |
|---|---|---|
| Perplexity | 4000 | 3400 |

| | SVI | 1 node Map Red | 1 node REEF | 4 node REEF |
|---|---|---|---|---|
| overall | 2 hours | 4 hours 31 mins | 68 mins | 36 mins |
| Whiten | | 16 mins | 16 mins | 16 mins |
| Matricize | | 15 mins | 15 mins | 4 mins |
| ALS | | 4 hours | 37 mins | 16 mins |