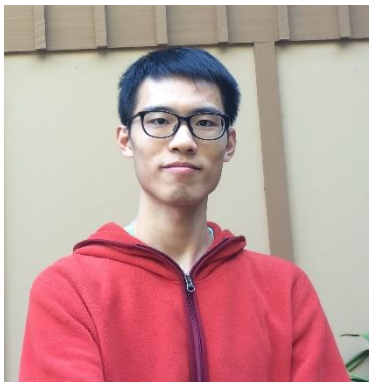
Smoothed Analysis for Tensor Decompositions and Unsupervised Learning

Aravindan Vijayaraghavan
Northwestern University

mostly based on joint work with

Aditya Bhaskara
University of Utah



Aidao Chen
Northwestern

Aidan Perreault
Northwestern -> Stanford

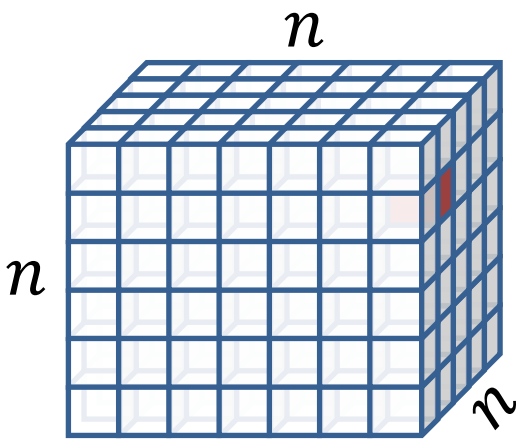
Outline

- Basic setting: definitions and CP decompositions
- Need for robustness (motivated by ML)
- Smoothed analysis for tensor decompositions
- Main results
- Overview of techniques (time permitting…)

Tensors and CP decompositions

Setting for this talk:

- Tensors of constant order ℓ e.g., $\ell = 3,4,6$ etc.
- n : dimension of each mode. n is large.
- Domain is reals. Tensors in $\mathbb{R}^{n \otimes \ell}$



Order $\ell = 3$

$T \in \mathbb{R}^{n \times n \times n}$

Tensor (CP) decompositions:

rank-r decomposition $T = \sum_{i=1}^{r} a_i \otimes b_i \otimes c_i$

Rank(T) = smallest such r possible. For order-ℓ , rank $r \leq n^{\ell-1}$

Symmetric rank-r decomposition $T = \sum_{i=1}^{r} a_i^{\otimes 3}$

- Useful property: uniqueness of tensor decompositions generically/ under mild conditions (for order-3 tensors and above).

Efficient algorithms for CP decomposition

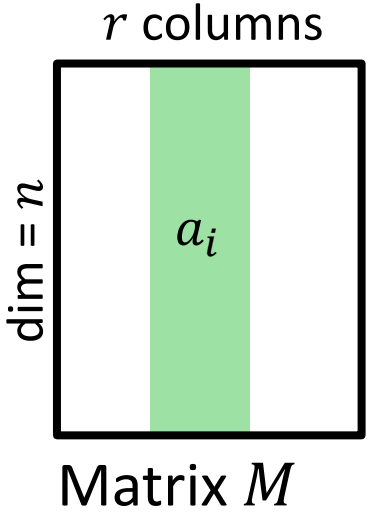
$$T = \sum_{i=1}^{r} a_i^{\otimes \ell}$$

(not w.l.o.g. for even ℓ)

Efficient algorithms: polynomial in input size $n^{O(\ell)}$ for order-ℓ i.e., $n^{c\ell}$ for some $c > 0$ (real arithmetic)

- NP-hard in worst case [Hillar, Lim'13] especially when $r \gg n$
- Efficient algorithms when $r \leq n$ (undercomplete setting)

Thm [Jennrich via Harshman'70] for $\ell = 3$ order.
Efficient algorithm that *recover* factors when the $\{a_1, \dots, a_r\}$ are linearly independent i.e., matrix M has rank r (hence $r \leq n$). (see Moitra's talk)



- Decomposition is unique even for larger rank $r > n$ (see e.g., [Kruskal'77, Chiantini-Ottaviani'07])

Efficient algorithms in the overcomplete setting i.e., rank $r \gg n$?
("beyond" worst-case settings)

Overcomplete setting with rank $r \gg n$?

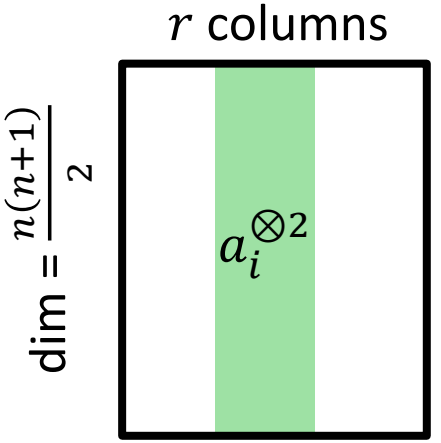
Yes, when order $\ell > 3$ *"generically"* $\quad T = \sum_{i=1}^{r} a_i^{\otimes \ell}$

- Generically: holds for all but a zero measure set of instances

$$T \quad (\ell = 6) = \sum_{i=1}^{r} \underbrace{a_i \otimes a_i}_{u_i} \otimes \underbrace{a_i \otimes a_i}_{u_i} \otimes \underbrace{a_i \otimes a_i}_{u_i} = \sum_{i=1}^{r} u_i \otimes u_i \otimes u_i$$

Jennrich's algorithm works if matrix M with ith column being $u_i = a_i^{\otimes 2}$ is linearly independent

- Holds for rank $r \leq n(n+1)/2$ generically
- For order ℓ, works generically $r \leq n^{\lfloor (\ell-1)/2 \rfloor}$



Matrix M for $\ell = 6$

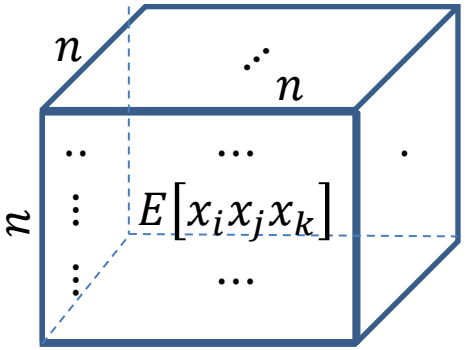
Thm [Cardoso'91, DeLathauwer, Castiang, Cardoso'07].
"FOOBI" algorithm for 4-tensors of rank r *generically* when $r \leq c.n^2$

Unsupervised Learning and the method of moments

Tensor decompositions for learning parameters of probabilistic model
e.g., mixtures of Gaussians, hidden Markov models, 2-layer neural nets

step 1. compute tensor (e.g., moments) whose CP decomposition *encodes* model parameters

step 2. find decomposition (and get parameters)



$$\boldsymbol{T = \sum_{i=1}^{k} f(\mu_i)} \qquad \text{e.g.}, T = \sum_{i=1}^{k} \mu_i^{\otimes 3}$$

where $f: \mathbb{R}^n \to \mathbb{R}^m$ is a degree-ℓ rank-1 polynomial in some of the parameters

Uniqueness $\Longrightarrow$ Recover parameters μ_i (identifiability)
Polynomial time algorithm for Decomposition $\Longrightarrow$ efficient learning

- Learning problems challenging in "*overcomplete*" settings ($k \gg n$)

Need for Robustness to Errors

In most applications, tensor estimated from samples (e.g., 3rd moment)

Beware : Sampling error and noise

With N samples, error per entry of tensor $\approx 1/\sqrt{\mathrm{N}}$

With $N = \mathrm{poly}(n)$ samples, get tensor T up to error $\epsilon = 1/\mathrm{poly}(n)$

$$T = \sum_{i=1}^{r} a_i^{\otimes \ell} + \mathrm{Err}$$

Note: T may have rank $\gg r$

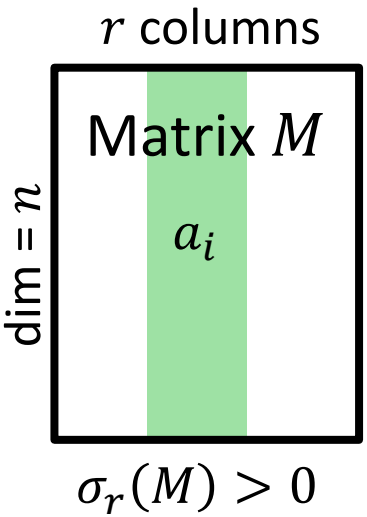
Low-rank ϵ-approximation:

Low-rank decomposition approximating T up to error ϵ in Frobenius norm i.e. $\left\| T - \sum_{i=1}^{r} a_i^{\otimes \ell} \right\|_F \leq \epsilon$

Goal: Algorithms taking time $\mathrm{poly}(\mathrm{n}^{\ell})$ robust to inverse-polynomial error in tensor T (Frobenius norm) e.g., recovery error is $\mathrm{poly}(\epsilon, \mathrm{n}^{\ell})$

Robust analogs of classic results

$T = \sum_{i=1}^{r} a_i^{\otimes 3} + E$, where $\|E\|_F \leq \epsilon$



Thm. Given tensor T as above with $\|a_i\|_2 \leq B$, and Jennrich's algorithm runs in polynomial time and w.h.p. recovers the factors a_i up to error δ (in ℓ_2) when

1. $\sigma_{min} = \sigma_r(M) \geq 1/poly(n)$ (linear independence).
2. $\|E\|_F = \epsilon \leq \eta/poly(n, B)$

- See e.g., [Goyal-Xiao-Vempala'14, Bhaskara-Charikar-Moitra-V'14, Moitra'16]
- Similar robust analog true for Kruskal's uniqueness theorem [BCV'14]

Robust analogs of results in overcomplete settings (rank $r \gg n$) ?

Robust guarantees: analog of genericity?

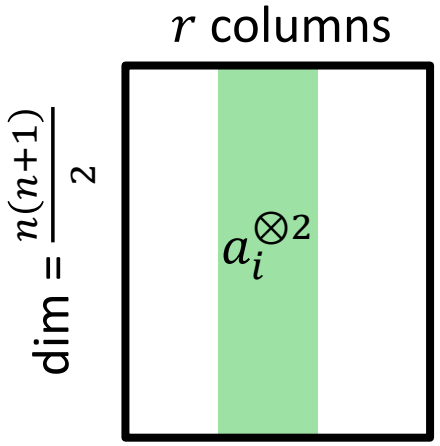
- Generically: holds for all but a zero measure set of instances

Jennrich's algorithm for order $\ell = 6$:

For robust guarantees, need $\sigma_r(M) > 1/poly(n)$

- For **M formed** $\{a_i^{\otimes 2} \mid i \in [r]\}$ as columns
- In general, lower bound on

$$\sigma_r(\ (a_i^{\otimes \ell} \mid i \in [r])\ )\ ?$$



matrix M for $\ell = 6$

$$\sigma_r(M) > \frac{1}{poly(n)}\ ?$$

Robust version of FOOBI [Cardoso'91, DlCC'07]:

Also need $\sigma_r(M) > 1/poly(n)$ for matrix M

- with columns : $\{a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2} \mid i, j \in [r], i < j\}$

How can we argue about such conditions holding most of the time?

Smoothed Analysis

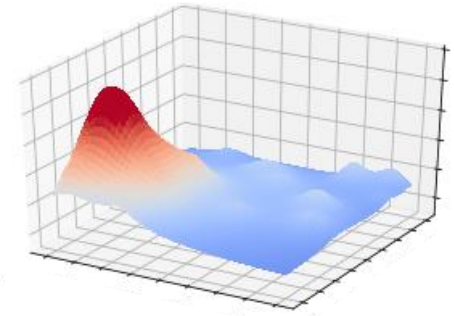
[Spielman & Teng 2000]

Simplex algorithm solves LPs efficiently (explains practice).

Smoothed analysis guarantees:

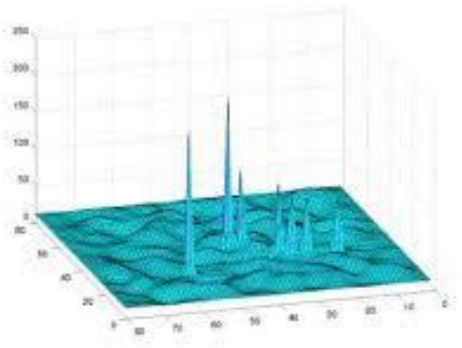
- Analyze small random perturbation of input makes instances easy
- Bad instances are isolated

- Smoothed analysis guarantees stronger qualitatively than average-case



Good Average case guarantees

vs



Smoothed analysis

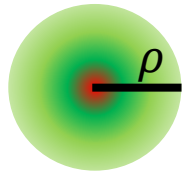
- A natural analog of “generic” results

Smoothed Analysis for Tensor Decompositions

Factors of the Decomposition are perturbed [BCM**V**'14]

1. Adversary chooses tensor $T = \sum_{i=1}^{r} a_i \otimes a_i \otimes \ldots \otimes a_i$

2. $\tilde{a}_i$ is random ρ-perturbation of a_i *i.e. add random vector drawn i.i.d. from Gaussian* $N(0, \frac{\rho^2}{n} I_{n\times n})$



$a_i \in \mathbb{R}^n \Rightarrow \tilde{a}_i$ expected length is ρ

think $\rho \sim \frac{1}{poly(n)}$

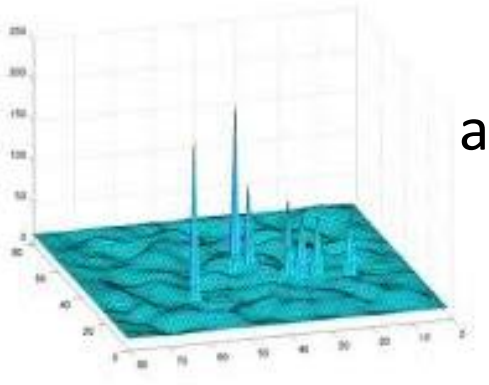
3. Input: $\tilde{T}$. Analyse algorithm on $\tilde{T}$.

$$\tilde{T} = \sum_{i=1}^{r} \tilde{a}_i \otimes \tilde{a}_i \otimes \ldots \otimes \tilde{a}_i + \text{noise}$$

Goal: Algorithm that finds the rank-r decomposition

- with high probability over random perturbation e.g., $1 - \exp(-cn)$
- run time, recovery error have polynomial dependence on $n, 1/\rho$

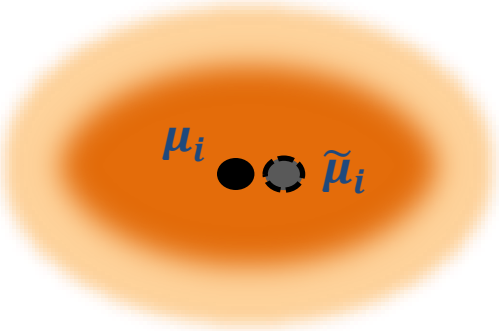
Smoothed Analysis for Learning



Smoothed setting: Parameters of the model are assumed to be random perturbed (by small amount).

Samples drawn from the perturbed model.

E.g., for mixtures of Gaussians, the means $\{\mu_i\}$ of the components are randomly perturbed slightly



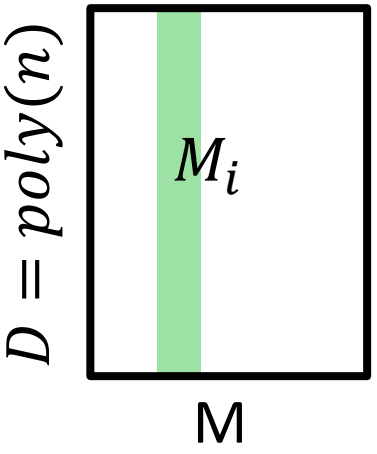
Smoothed analysis for Tensor methods

Often need to prove $\sigma_{min}(M) \geq 1/poly(n)$ w.h.p. for M given by

Overcomplete Tensor Decompositions:

- **Jennrich algorithm:** M has columns $\{a_i^{\otimes \ell} \mid i \in [r]\}$
- **FOOBI algorithm:** M has columns :

$$\{a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2} \mid i, j \in [r], i < j\}$$

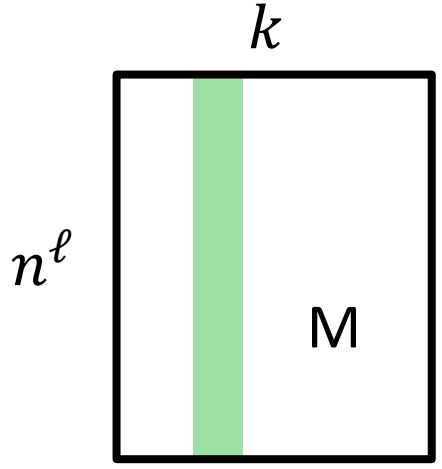


Implementing method of moments (learning):

- **Mixtures of spherical Gaussians:** M has columns $\left\{\tilde{\mu}_i^{\otimes \ell} \mid i \in [r]\right\}$ where $\{\tilde{\mu}_i\}$ are the unknown mean parameters
- **Hidden Markov Models:** M has columns $\left\{\tilde{a}_{i_1} \otimes \tilde{a}_{i_2} \otimes \cdots \otimes \tilde{a}_{i_\ell}\right\}$, where $(i_1, \ldots, i_\ell)$ is a path through state space, and $\{\tilde{a}_i\}$ parameters

For different models, perturbation of parameters $\Longrightarrow$ different kinds of perturbations to terms of decomposition

Towards a general framework?



Every entry of M is a low-degree polynomial of underlying factors $a_1, a_2, a_3 \ldots \in \mathbb{R}^n$:
Random matrix ensembles with highly dependent entries.

- High-confidence lower bounds on least singular value imply polynomial time guarantees (w.h.p).
- Degenerate instances ⇔ matrix M is singular

(nice characterizations from algebraic geometry for singularity of M)

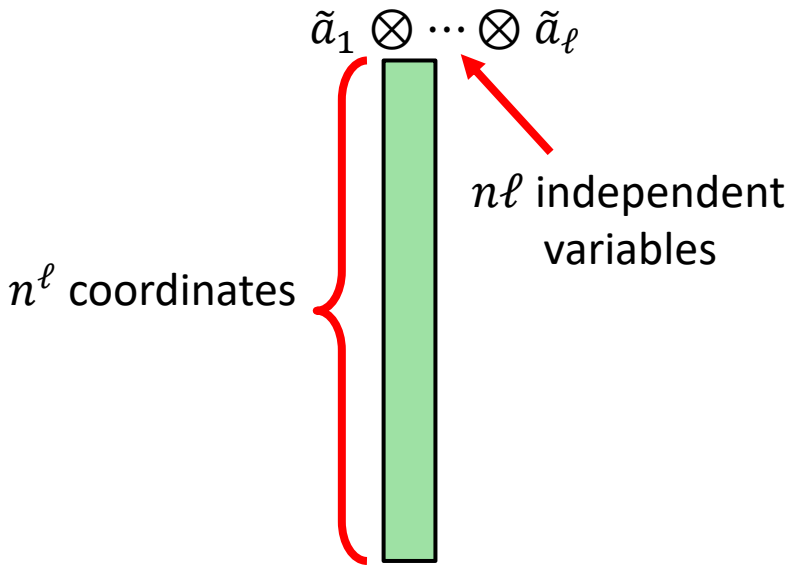
- Related to upper bound on the condition number of M
- Many problems in unsupervised learning fit this framework

Goal: Give inverse polynomial lower-bounds on $\sigma_{min}(M)$ for such general random matrix ensembles (with dependent entries)

Challenges towards a general framework?

Goal: Give inverse polynomial lower-bounds on $\sigma_{min}(M)$ for M that has columns given by some polynomials in a few $\tilde{a}_i$?

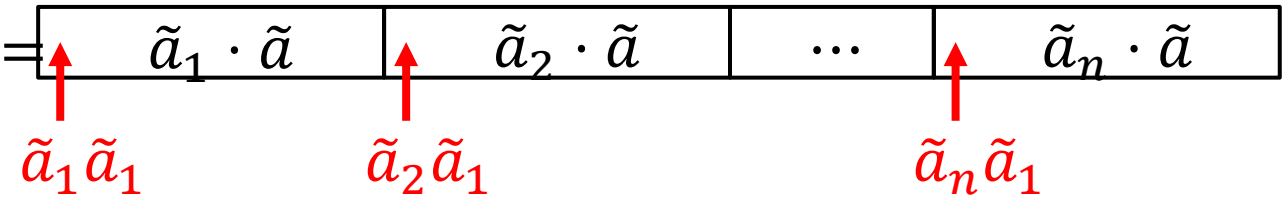
- Lower bounding $\sigma_{min}(M)$ is difficult even for random matrix with fully independent entries (random matrix theory)
- Lots of dependencies.



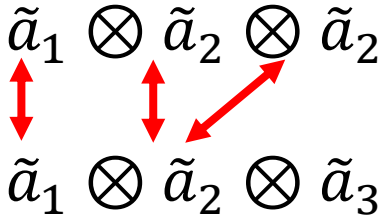
Challenges: lots of dependencies

Within one column of M:

$f(\tilde{a}) = \tilde{a}^{\otimes 2} =$



Between columns:



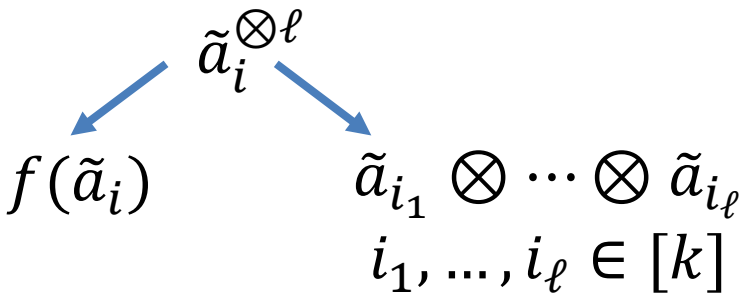
OUR RESULTS

Results

Goal: Give inverse poly(n) lower-bounds on $\sigma_{min}(M)$ for random matrix ensembles (dependent entries) given by polynomials

Bounds on $\sigma_{min}(M)$ in two general settings:

- Each column M_i of M is a vector-valued homogeneous polynomial in $\tilde{a}_i$.
- Each column of M is a tensor product of a few $\tilde{a}_i$.

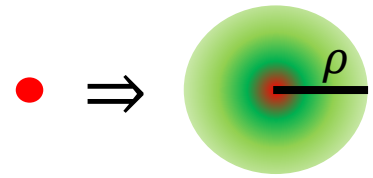


Applications to unsupervised learning problems:

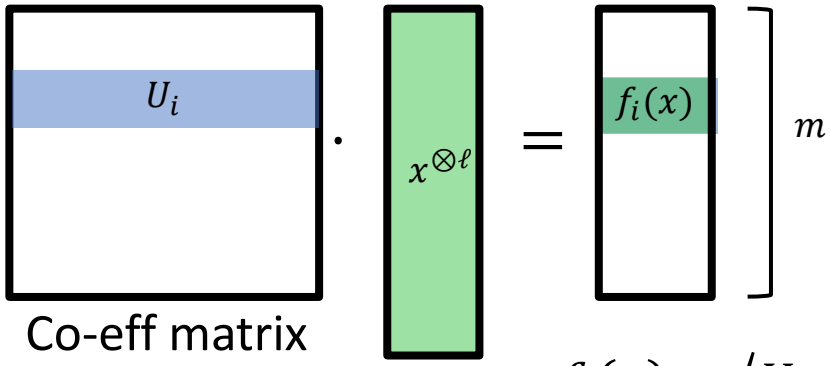
- Higher order tensor decomposition
- Gaussian mixture models
- Hidden Markov Models
- 2-layer neural nets
- Robust subspace recovery etc..

Polynomials of one perturbed vector

Each column is a fixed polynomial applied to a different perturbed vector

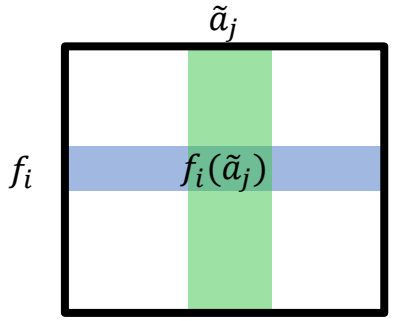


$a_j \in \mathbb{R}^n \Rightarrow \tilde{a}_j$



Co-eff matrix

vector in $\mathbb{R}^{n^\ell}$ $f_i(x) = \langle U_i, x^{\otimes \ell} \rangle$



$M_f \in \mathbb{R}^{m \times r}$

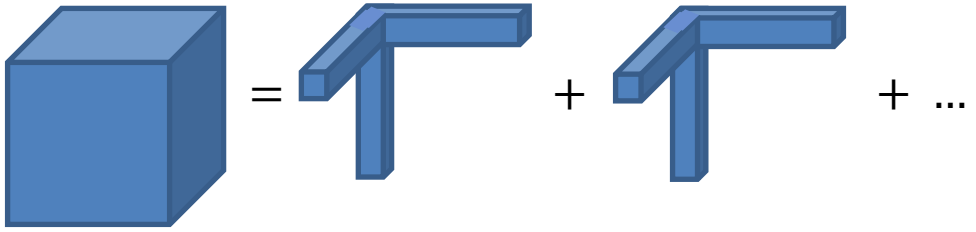
Theorem 1: With probability at least $1 - \exp(-\Omega_\ell(\delta n) + \log r)$,

$$\sigma_{min}(M_f) \gtrsim \frac{\rho^\ell}{n^\ell} \cdot \sigma_{r+\delta n^\ell}(U).$$

- Condition on U measures how different $f_1, \ldots, f_m$ are.
- Result is almost tight:
  - Failure probability
  - Number of large singular values of U.

Consequences for symmetric tensor decompositions using Jennrich [BCMV'14]

$$T = \sum_{i=1}^{r} \tilde{a}_i^{\otimes \ell} + \mathrm{Err}$$

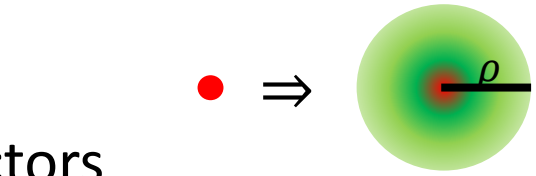


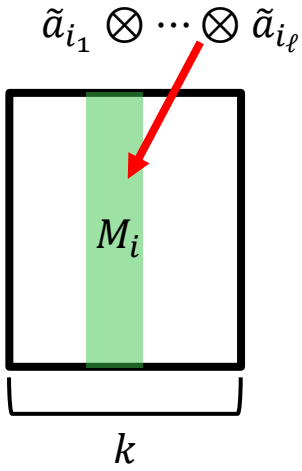
Thm. If $r \lesssim n^{\lfloor \frac{\ell-1}{2} \rfloor}$, algorithm runs in polynomial time and recovers $\{\tilde{a}_i : i \in [r]\}$ from T up to error ϵ with probability at least $1 - \exp(-\Omega_\ell(n))$, as long as $\|\mathrm{Err}\|_F \leq 1/poly(n, \frac{1}{\epsilon}, \rho, \max_i \|a_i\|)$.

- Polynomial time (smoothed analysis) learning guarantees for mixtures of k-spherical Gaussians when $k \leq n^\ell$ for any $\ell > 0$.
- Uses Theorem 1 which $f(x) = x^{\otimes \lfloor (\ell-1)/2 \rfloor}$
- Similar guarantees (with slightly weaker parameters) shown by Bhaskara-Charikar-Moitra-V'14 (using techniques specific to this).

Polynomials of a few perturbed vectors

Each column is a monomial of a few perturbed vectors





Overlap measure Δ_s :
For each $s \in [\ell]$: Δ_s is an upper bound on number of other columns that disagree from column i in exactly s terms of the monomial

$\tilde{a}_1 \otimes \tilde{a}_2 \otimes \tilde{a}_3 \otimes \tilde{a}_4$
$\tilde{a}_1 \otimes \tilde{a}_2 \otimes \tilde{a}_3 \otimes \tilde{a}_5$

$\ell - 1$ overlaps
$\Rightarrow$
only $O(n)$ dimensions

$\tilde{a}_1 \otimes \tilde{a}_2 \otimes \tilde{a}_3 \otimes \tilde{a}_4$
$\tilde{a}_1 \otimes \tilde{a}_5 \otimes \tilde{a}_6 \otimes \tilde{a}_7$

1 overlap
$\Rightarrow$
$O(n^{\ell-1})$ dimensions

- Overlap condition is tight

Overlap condition: $\sum_{s=1}^{\ell} \frac{\Delta_s}{n^s} \leq O_\ell(1)$

Theorem 2: If the columns of M satisfy this few "overlaps condition", then with probability at least $1 - \exp(-\Omega_\ell(\delta n) + \log k)$,

$$\sigma_{min}(M) \gtrsim \frac{\rho^\ell}{n^\ell}.$$

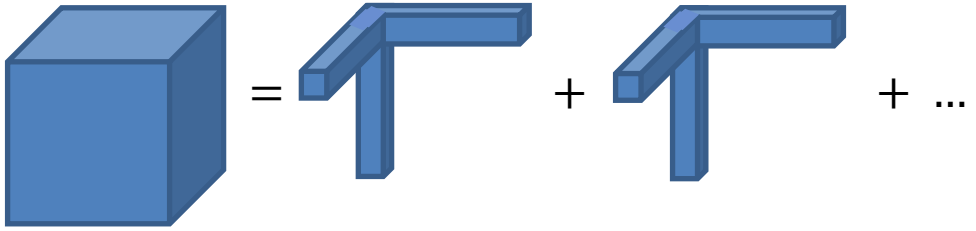
Previous work

Our work builds on prior work that handles special cases (tailor-made for the specific application):

- "Decoupled multilinear case": If M has columns $M_i = \tilde{a}_i^{(1)} \otimes \cdots \otimes \tilde{a}_i^{(\ell)}$ [Bhaskara Charikar Moitra V '14]. See also [Anderson, Belkin, Goyal, Rademacher, Voss '14] for some weaker bounds.
- [Anari, Daskalakis, Maass, Papadimitriou, Saberi, Vempala '18] improved dependencies and generalized it to more distributions (e.g., discrete distributions).
- Other extensions/ special cases related to tensor decompositions [Ma, Shi, Steurer '16]

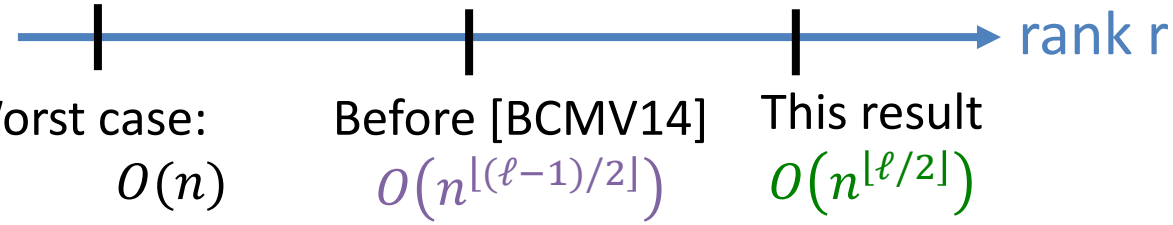
Robust tensor decomposition

$$T = \sum_{i=1}^{r} \tilde{a}_i^{\otimes \ell} + \mathrm{Err}$$



Thm. Let ℓ be even. If $r \lesssim n^{\ell/2}$, algorithm given T runs in polynomial time and recovers $\{\tilde{a}_i : i \in [r]\}$ up to error ϵ with prob. at least $1 - \exp(-\Omega_\ell(n))$, if $\|\mathrm{Err}\|_F \leq 1/\mathrm{poly}(n, \frac{1}{\epsilon}, \rho, \max_i \|a_i\|)$.

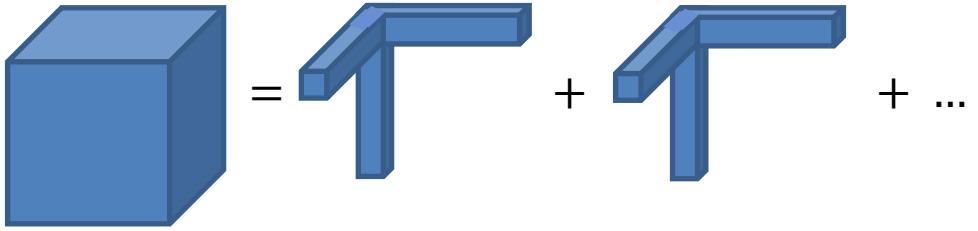
- New algorithm based on a generalization of FOOBI algorithm [Cardoso '91], [De Lathauwer, Castaing, Cardoso '07]



e.g., when $\ell = 6$, improves from rank $O(n^2)$ to $O(n^3)$

FOOBI Algorithm and robustness

$$T = \sum_{i=1}^{r} \tilde{a}_i^{\otimes 4} + \mathrm{Err}$$



- Based on constructing a "rank-1" detecting device
- Find eigenvectors of $n^2 \times n^2$ matrix from flattening to find the span of rank-1 $n \times n$ matrices
- Set up a system of equations using rank-1 detecting device, where coefficients involve eigenvectors

Challenge with robustness: Linear system coefficients are non-linear functions of eigenvectors, which is more brittle to noise.

- Use Theorem 2 along with careful robustness analysis
- See also [Ma-Shi-Steurer'16], [Hopkins-Shi-Schramm'20] for robust analysis of a different algorithm for $\ell = 4$.
- Also generalize and give robust analysis for higher even ℓ.

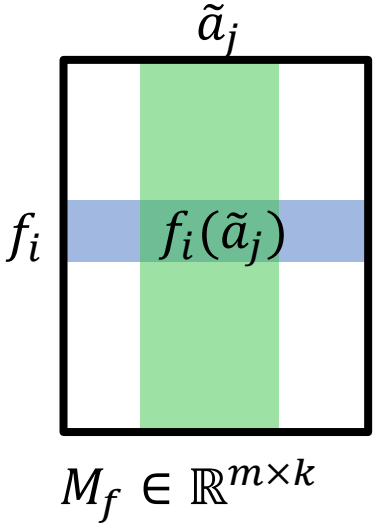
PROOF TECHNIQUES

Proof of Theorem 1

Theorem 1: With probability at least $1 - \exp(-\Omega_\ell(\delta n) + \log k)$,

$$\sigma_{min}(M_f) \gtrsim \frac{\rho^\ell}{n^\ell} \cdot \sigma_{k+\delta n^\ell}(U).$$

$f(x) = U \cdot x^{\otimes \ell}$



Leave-one-out distance: min. distance of each column from the span of the rest of columns

$$L(M) \coloneqq \min_i \left\| \Pi_{M_{-i}}(M_i) \right\|_2.$$

Fact: $\frac{L(M)}{\sqrt{k}} \leq \sigma_{min}(M) \leq L(M)$

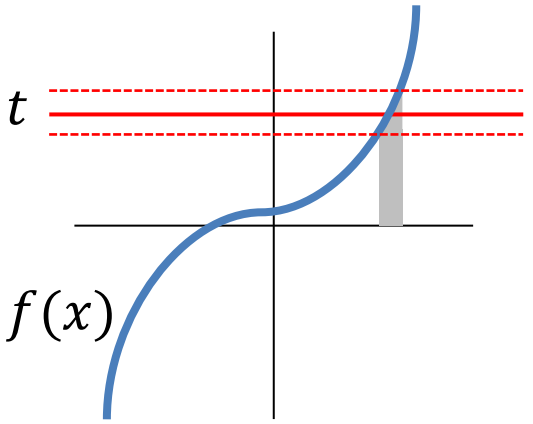
Goal: Show that kth column is far from span of the other columns

Relation to anti-concentration of polynomials

Anti-concentration inequality [Carbery, Wright '01, NSV'03]

Let $x \sim N(0,1)^n$, $f: \mathbb{R}^n \to \mathbb{R}$ a degree ℓ polynomial with $\mathrm{Var}\big(f(x)\big) = 1$.

$$\mathbb{P}_x\left[|f(x) - t| < \epsilon\right] \leq O\left(\epsilon^{1/\ell}\right).$$



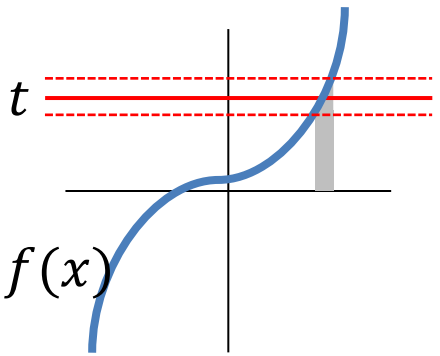
Problem: Failure probability is too weak for good guarantees (it has polynomial dependence on σ_{min}, and algorithm performance typically depends as poly$(1/\sigma_{min})$)

Goal: Need super polynomial small failure probability (e.g., exponential small probability)

Generalization of Carbery-Wright to vector-valued functions

Carbery-Wright inequality [CW'01]. Let $f: \mathbb{R}^n \to \mathbb{R}$ be a degree ℓ polynomial with $\mathrm{Var}\big(f(x)\big) = 1$.

$$\mathbb{P}_{x\sim N(0,1)^n}\left[|f(x) - t| < \epsilon\right] \le O\big(\epsilon^{1/\ell}\big).$$



Consider $f = (f_1, \dots, f_m)$: $\mathbb{R}^n \to \mathbb{R}^m$ of degree ℓ .

Qn: If $f_1, \dots, f_m$ were ``independent'' e.g., m linear functions given by orthonormal vectors, what would the failure probability be?

$$\mathbb{P}_x\left[\,\|f(x) - t\|_\infty < \epsilon\,\right] = \mathbb{P}_x\left[\,\forall j \in [m], \left|f_j(x) - t\right| < \epsilon\,\right] \le O\big(\epsilon^{m/\ell}\big)$$

Proposition: (Thm. 1 with $k = 1$) Let $f = U\, x^{\otimes \ell}$. If $\sigma_{\delta n^\ell}(U) > \eta$,

$$\mathbb{P}\left[\|f(\tilde{x}) - t\|_2 < O(\epsilon\eta)\cdot\frac{\rho^\ell}{n^\ell}\right] < \epsilon^{c_\ell \cdot \delta n}.$$

- $\sigma_{\delta n^\ell}(U)$ of the co-efficient matrix U measures how different $f_1, \dots, f_m$ are.

The main steps in the proof

Goal: Show that $\left\| \Pi_W \tilde{a}_i^{\otimes \ell} \right\|_2$ is large for any subspace $W \subset \mathbb{R}^{n^\ell}$ of dimension at least $0.1\, n^\ell$ w.h.p.

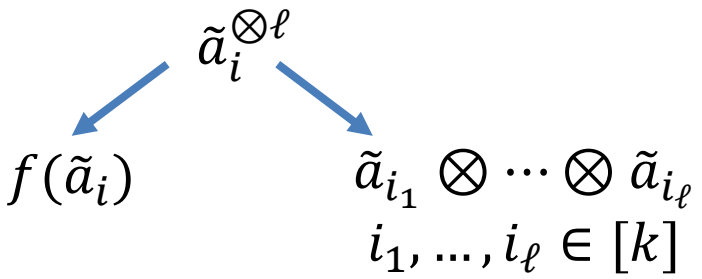
1. "Decoupling" ideas from probability to reduce the above case to the simpler setting where every column is a tensor product of different, independent random vectors.
2. Use known bounds for the "decoupled" multilinear case [BCMV'14, ADMPSV'18]

Summary of Results

Goal: Give inverse poly(n) lower-bounds on $\sigma_{min}(M)$ for random matrix ensembles (dependent entries) given by polynomials

Bounds on $\sigma_{min}(M)$ in two general settings:

- Each column M_i of M is a vector-valued homogeneous polynomial in $\tilde{a}_i$.
- Each column of M is a tensor product of a few $\tilde{a}_i$.



Applications to unsupervised learning problems:

- Robust subspace recovery
- Hidden Markov Models
- Higher order tensor decomposition

Future Directions

Robust analogs of other "generic" results

- Uniqueness for tensor CP decompositions for rank $r \leq cn^2$ generically e.g., [Chiantini-Ottaviani'12]

More general random matrix ensembles

- Is there some hope of a more complete characterization as in algebraic geometry?

Smoothed Analysis for other Learning problems ?

Thank You! Questions?

Based on works…

- **[BCPV'19]** *Smoothed Analysis in Unsupervised Learning via Decoupling*. Bhaskara, Chen , Perreault and Vijayaraghavan. *FOCS 2019,* and

  Smoothed Analysis for Tensor Methods in Unsupervised Learning. Mathematical Programming Series B, 2020.

- **[BCPV'14]** *Smoothed Analysis of Tensor Decompositions.* Bhaskara, Charikar , Moitra and Vijayaraghavan. *STOC 2014.*

- [Book Chapter] *Efficient Tensor Decomposition.* Vijayaraghavan. Book Chapter in *Beyond the Worst Case Analysis of Algorithms,* edited by Tim Roughgarden.