Computational lower bounds for Tensor PCA

Daniel Hsu (Columbia University)
Based on joint work with Rishabh Dudeja (Harvard University)

UCSD Theory Seminar
April 17, 2023

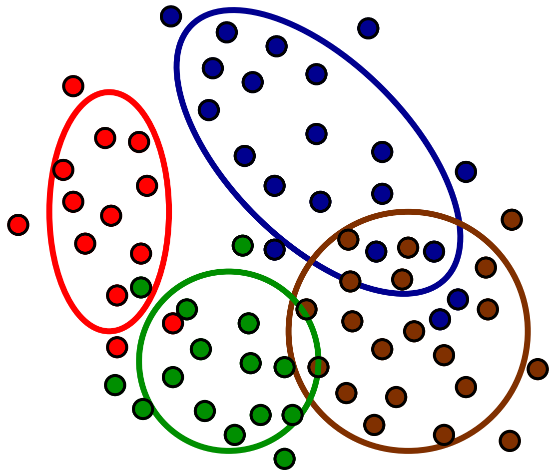
Outline

- **Main result**: "Memory size × Sample size × Time" lower bounds for Tensor PCA and related problems
- **Talk outline**:
    1. Motivation from statistical modeling
    2. Tensor PCA and our lower bounds
    3. Memory-bounded algorithms for Tensor PCA
    4. High-level proof ideas

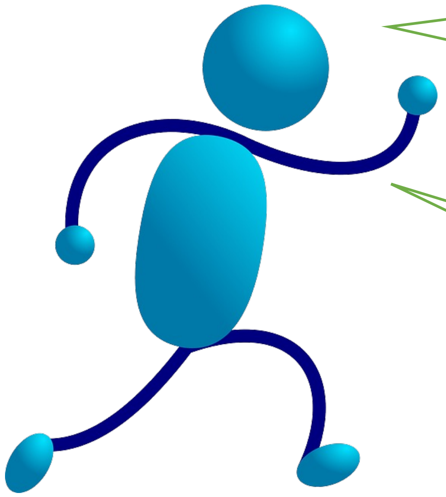
1. Motivation

Fitting statistical models to multivariate data

- **Statistical model**: e.g., mixture of Gaussians
$$Y_1, \dots, Y_n \sim_{\mathrm{iid}} w_1 \mathrm{N}(\mu_1, \Sigma_1) + w_2 \mathrm{N}(\mu_2, \Sigma_2) + \cdots$$
- **Model fitting**: Find model parameters $(w_1, \mu_1, \Sigma_1, w_2, \mu_2, \Sigma_2, \dots)$ of probability distribution that "best fits the data" $y_1, \dots, y_n \in \mathbb{R}^d$



How to estimate parameters?



How should I choose the model parameters to fit my multivariate data?

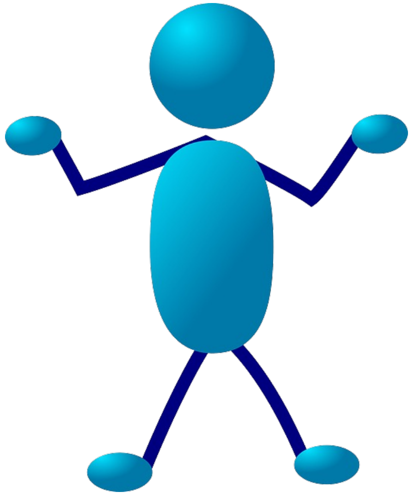
Maximum Likelihood Estimation (MLE)!

But likelihood is NP-hard to optimize …
[Tosh & Dasgupta, '18; …]

That's in the worst case. Your data may be nicer …

You're right---local search works well sometimes!
[Dasgupta & Schulman, '07; Xu, H., Maleki, '16; …]

Oops, local search can fail even on "best case" data.
[Jin, Zhang, Balakrishnan, Wainwright, Jordan, '16]



Method of moments

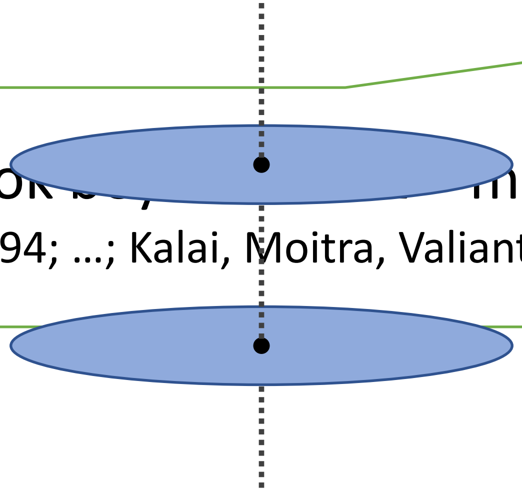
Have you tried the "method of moments"?

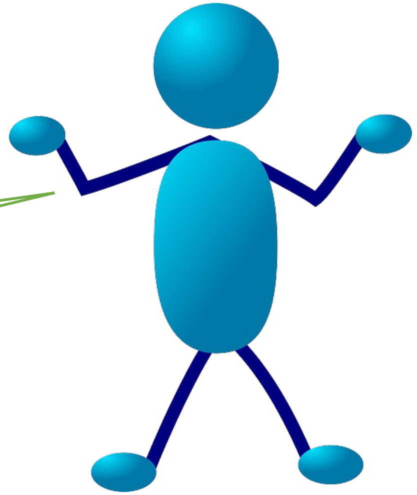
Like PCA? Yes, but it can be uninformative!

[Achlioptas & McSherry, '05]

You can loo[illegible] moment …

[Pearson, '94; …; Kalai, Moitra, Valiant, '10; …]





Spherical Gaussians [Vempala & Wang, '02]

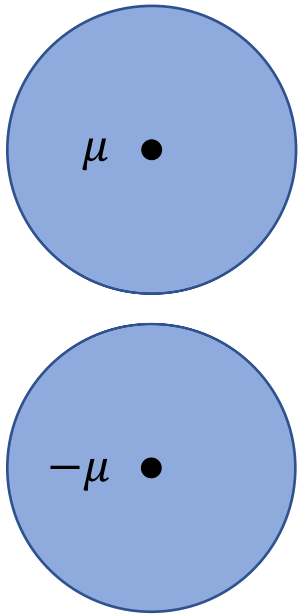
$$Y \sim \frac{1}{2}\mathrm{N}(-\mu, I_d) + \frac{1}{2}\mathrm{N}(\mu, I_d)$$

2nd moment matrix reveals μ

$$\mathbb{E}[YY^\top] = I_d + \mu\mu^\top$$

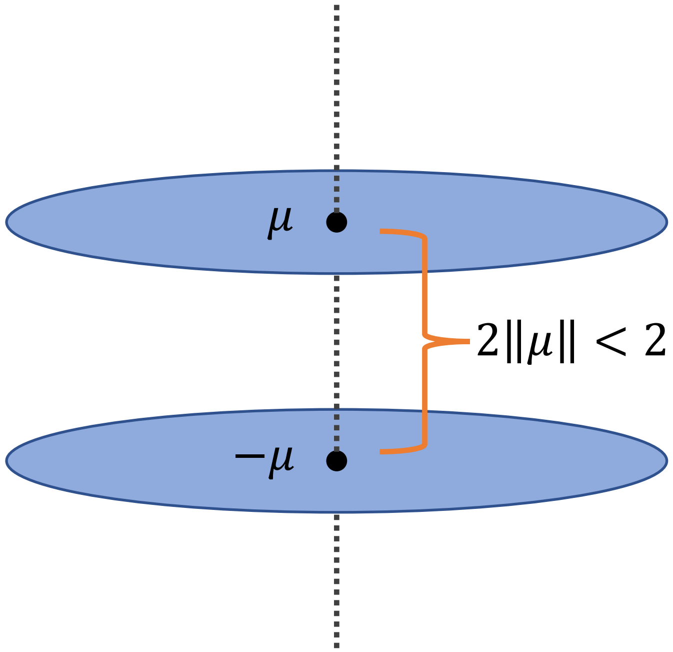
Top eigenvector of $\mathbb{E}[YY^\top]$ is $\propto \mu$

"Principal Components Analysis (PCA)"



Parallel Pancakes

$$Y \sim \frac{1}{2}\mathrm{N}(-\mu, I_d - \mu\mu^\top) + \frac{1}{2}\mathrm{N}(\mu, I_d - \mu\mu^\top)$$



2nd moment matrix is not useful

$$\mathbb{E}[YY^\top] = I_d$$

But 4th moment tensor reveals μ

$$\mathbb{E}\left[Y^{\otimes 4}\right] - \mathrm{Sym}(I_d \otimes I_d) = -\frac{1}{8}\mu^{\otimes 4}$$

Problem: All known poly-time algorithms for estimating μ this way require $n \gtrsim d^2$, even though MLE only needs $n \gtrsim d$

Does computational tractability come with a statistical cost?

2. Tensor PCA

Tensor PCA [Montanari & Richard, '14]

Asymmetric Tensor PCA:

$$X_i = \lambda\, \theta_1 \otimes \theta_2 \otimes \cdots \otimes \theta_k + Z_i$$

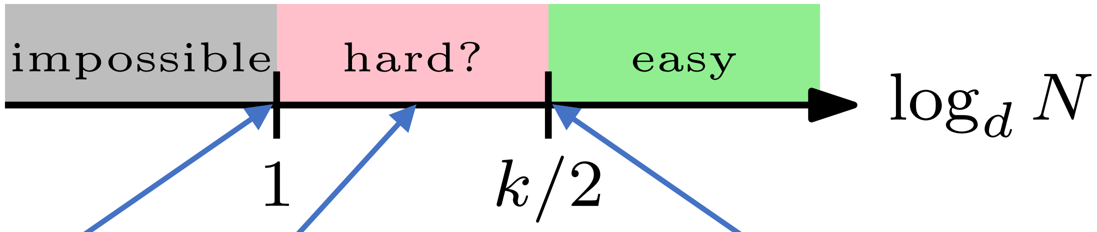
- $\theta_1, \dots, \theta_k \in \Theta \subseteq S^{d-1}$: k parameter vectors
- $(i_1, i_2, \dots, i_k)$ entry of $\theta_1 \otimes \theta_2 \otimes \cdots \otimes \theta_k$ is $\theta_1(i_1)\theta_2(i_2)\cdots\theta_k(i_k)$

- **Data model**: iid random order-k tensors $X_1, \dots, X_N$ in $\otimes^k \mathbb{R}^d$

$$X_i = \lambda\, \theta^{\otimes k} + Z_i$$

  - $\theta \in \Theta \subseteq S^{d-1}$: parameter vector to estimate (up to sign) within ℓ_2 error 0.01
  - $\lambda^2 > 0$: signal-to-noise ratio per data point
  - Z_i: order-k tensor of d^k iid $\mathrm{N}(0,1)$ random variables
  - $(i_1, i_2, \dots, i_k)$ entry of $\theta^{\otimes k}$ is $\theta(i_1)\theta(i_2)\cdots\theta(i_k)$
- **Motivations**:
  - $k = 2$: model problem for studying PCA ("spiked Wigner model")
  - $k \geq 3$: model problem for studying tensor-based method-of-moments
- *Sample complexity? Computational complexity?*

Statistical-to-computational gap



Information-theoretic lower bound:
No algorithm works with $N \lesssim d/\lambda^2$

Known poly-time algorithms:
[MR'14, HSS'15, ZT'15, HSSS'16, …]
Require $N \gtrsim d^{k/2}/\lambda^2$

$k \geq 3$: **Reasons to believe hardness?**

- Failure of specific poly-time algorithms [MR'14, BAGJ'20, HKPRSS'17]
- Hypergraphic Planted Clique [ZX'18; BB'20]
- Hard in SQ model [DH'21; BBHLS'21]

$k = 2$: Data $X_1, \ldots, X_N$ are matrices
Solution: Find top eigenvector/singular vectors

- $\log d$ iterations of **power method**
- Just need $N \gtrsim d/\lambda^2$
- No gap between impossible & easy regimes!

Our results

- We show that existing poly-time algorithms for Tensor PCA are on Pareto frontier in terms of run-time, sample size, and memory size

- **Theorem** [Dudeja & H., 2022]: Every algorithm for TPCA(d, k, λ^2) that accurately estimates the parameters must use

$$\text{memory size} \times \text{sample size} \times \text{time} \gtrsim \frac{d^{\lceil (k+1)/2 \rceil}}{\lambda^2}$$

  - For Asymmetric Tensor PCA, get lower bound of d^k/λ^2
  - Similar results for related problems, including "Parallel Pancakes"
  - Current best poly-time algorithms match these lower bounds

3. Memory-bounded algorithms

Memory-bounded algorithms

Template for ($\boldsymbol{B}, \boldsymbol{N}, \boldsymbol{T}$) algorithm

- Initialize memory state $\in \{0,1\}^B$
- For iteration $t = 1,2,\ldots,T$:
  - For data point $i = 1,2,\ldots,N$:
    - $\text{state} \leftarrow \text{update}_{t,i}(\text{state}, X_i)$
- Return $\hat{\theta}(\text{state})$

Example: MLE via exhaustive search

$$\text{argmax}_{\hat{\theta}\in\Theta}\langle \bar{X}, \hat{\theta}^{\otimes k}\rangle$$

where $\bar{X} = (X_1 + \cdots + X_N)/N$

- $\langle \bar{X}, \hat{\theta}^{\otimes k}\rangle = \sum_{i=1}^{N}\langle X_i/N, \hat{\theta}^{\otimes k}\rangle$
- For fixed $\hat{\theta}$, can compute sum in single pass over data ($= 1$ "iteration")
- State tracks best obj. value and best $\hat{\theta}$
- Memory size required: $B = O(d)$
- Sample size required: $N = O(d)$
- Iterations: $T = 2^d$ $\quad (\Theta = \{\pm 1/\sqrt{d}\}^d)$

Algorithm for Asymmetric Tensor PCA (k=4)

Matricization algorithm [MR'14]

- Let $A = \mathrm{reshape}(\bar{X}) \in \mathbb{R}^{d^2 \times d^2}$
- $(\hat{u}, \hat{v}) =$ top singular vectors of A
- $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4) =$ (something w/ $\hat{u}, \hat{v}$)

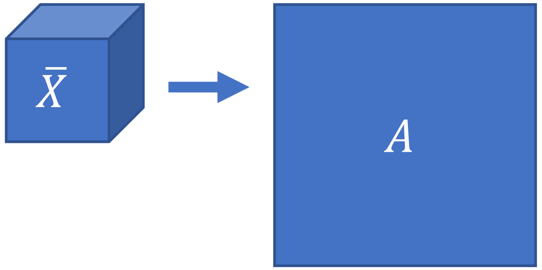
Recall:

$$\bar{X} \sim \lambda\, \theta_1 \otimes \theta_2 \otimes \theta_3 \otimes \theta_4 + \frac{1}{\sqrt{N}} Z$$

Matricization:

$$A \sim \lambda\, u \otimes v + \frac{1}{\sqrt{N}} \mathrm{reshape}(Z)$$

$$u = \mathrm{vec}(\theta_1 \otimes \theta_2),\ v = \mathrm{vec}(\theta_3 \otimes \theta_4)$$



$\mathrm{ATPCA}(d, 4, \lambda^2) \rightarrow \mathrm{ATPCA}(d^2, 2, \lambda^2)$

Sample size requirement

$$N \asymp \frac{d^2}{\lambda^2}$$

Power method impl.

Memory size: $B \asymp d^2$

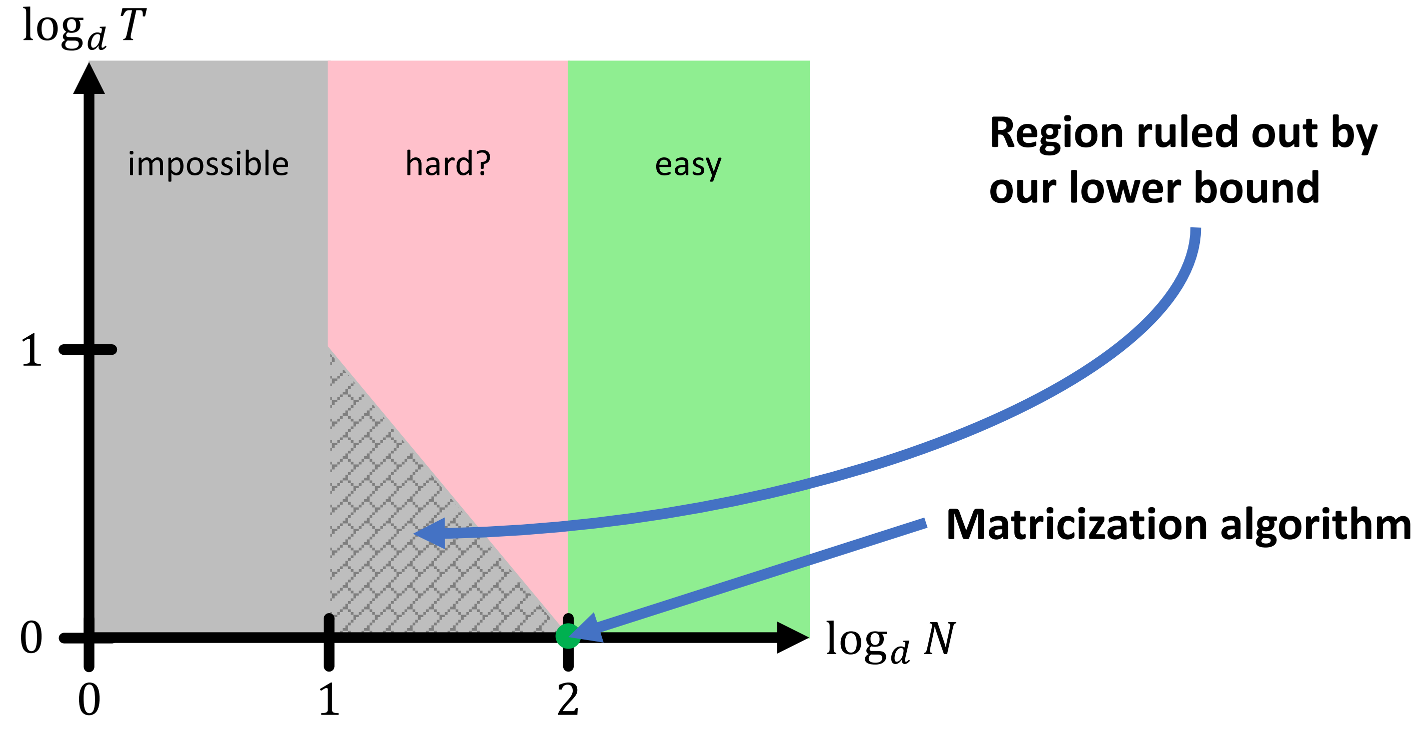
Iterations: $T \asymp \log d$

Total resources

$$BNT \asymp \frac{d^4}{\lambda^2} \log d$$

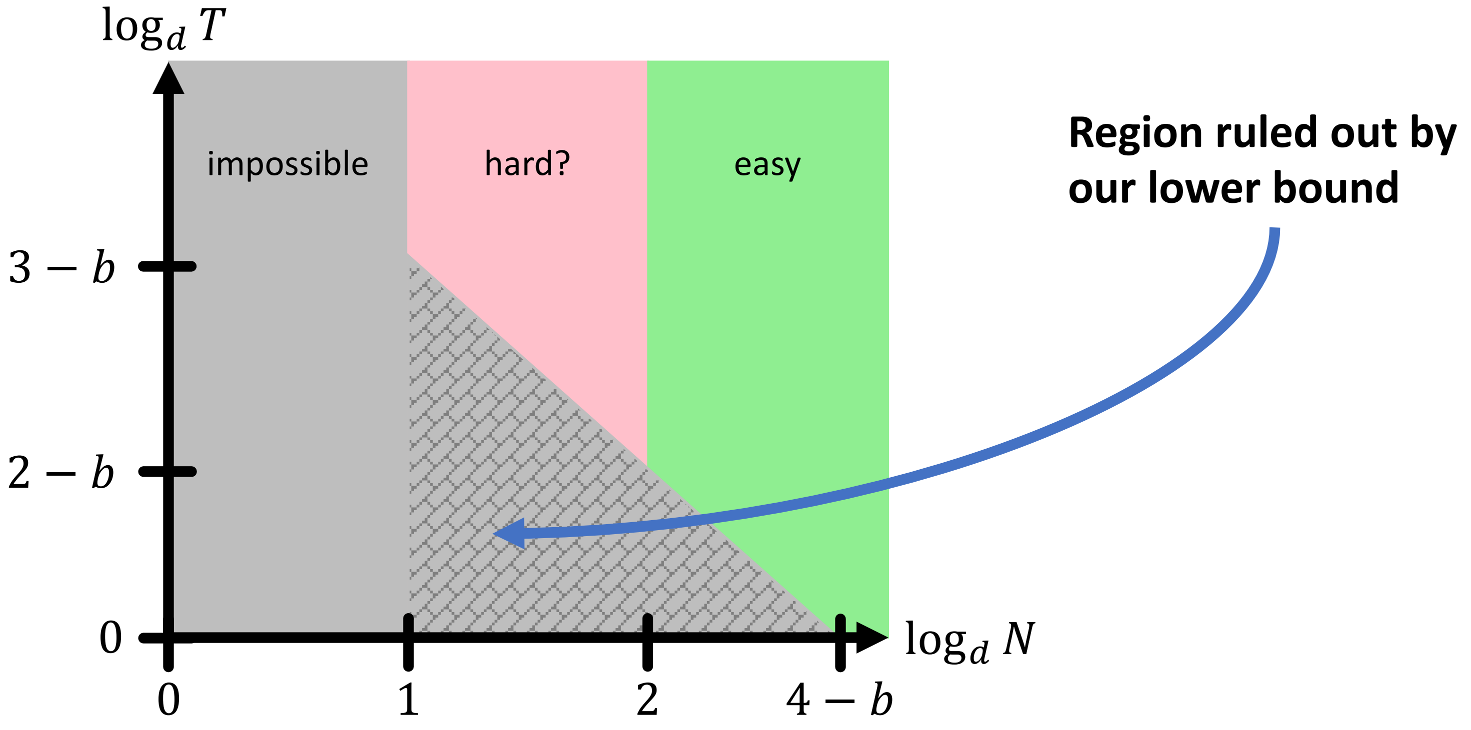
Phase diagram for Asymmetric Tensor PCA

- "Overparameterized" algorithms with $B \asymp d^2$



Need for overparameterization in ATPCA

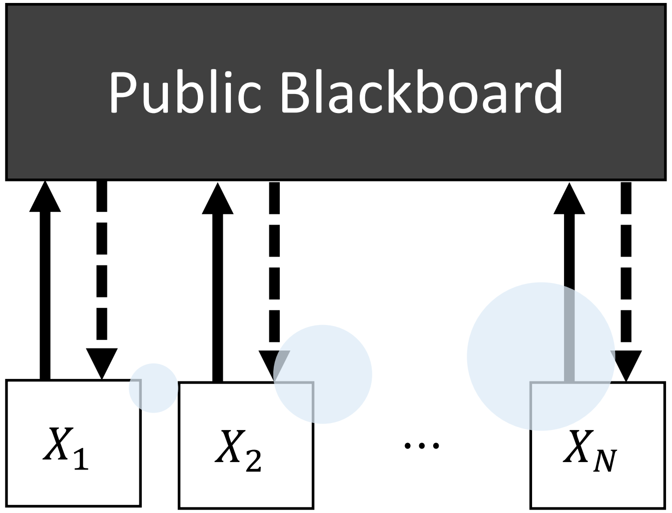
- Insufficiently overparameterized algorithms with $B \asymp d^b$ and $b < 2$



4. Proof ideas

Proof strategy: communication complexity

- Reduction from distributed estimation in blackboard model
  [Shamir, '14; Dagan & Shamir, '18]
  - (B, N, T) algorithm → protocol where each of N machines writes BT bits
- We prove new communication lower bounds for Tensor PCA



(B, N, T) algorithm

- Initialize memory state $\in \{0,1\}^B$
- For iteration $t = 1,2,\ldots,T$:
  - For data point $i = 1,2,\ldots,N$:
    - $\text{state} \leftarrow \text{update}_{t,i}(\text{state}, X_i)$
- Return $\hat{\theta}(\text{state})$

Lower bound via Fano's inequality

- Key quantity: Hellinger information [Chen, Guntuboyina, Zhang, '16]

$$I_{\mathrm{h}}(\theta; Y) = \inf_{Q} \int \mathrm{h}^2(P_\theta; Q)\pi(\mathrm{d}\theta)$$

  - $\mathrm{h}^2(\cdot;\cdot)$ is squared Hellinger distance
  - π is a prior distribution for parameter θ
  - P_θ is distribution of protocol transcript Y given θ
- If $I_{\mathrm{h}}(\theta; Y) \to 0$ as $d \to \infty$, then for large enough d, every protocol fails in average case sense with $\theta \sim \pi$ (and hence also for worst θ)
- We prove $I_{\mathrm{h}}(\theta; Y) \to 0$ if total communication $\ll \frac{d^{\lceil (k+1)/2 \rceil}}{\lambda^2}$ bits

Hellinger information bound

- New Hellinger information bound (simplified):

$$I_{\mathrm{h}}(\theta; Y) \lesssim \sum_{i=1}^{N} \mathbb{E}_0 \left[\int \left(\mathbb{E}_0 \left[\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu_0}(X_i) - 1 \middle| Y \right] \right)^2 \pi(\mathrm{d}\theta) \right]$$

  - $\mathbb{E}_0$ regards $X_1, \ldots, X_N$ as iid from null distribution μ_0
  - μ_θ is sampling distribution with parameter $\theta \in \Theta$
- What info does transcript Y have about (centered) likelihood ratios?

$$\left(\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu_0}(X_i) - 1 \right)_{\theta \in \Theta}$$

  - Need to bound squared "2-norm" of centered likelihood ratio process

Linearization and concentration

- Linearization of "2-norm" $\|v\|_\pi = \sqrt{\int v(\theta)^2 \pi(\mathrm{d}\theta)}$:

$$\left\| \mathbb{E}_0 \left[\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu_0}(X_i) - 1 \middle| Y \right] \right\|_\pi = \sup_{\|v\|_\pi = 1} \left\langle v, \mathbb{E}_0 \left[\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu_0}(X_i) - 1 \middle| Y \right] \right\rangle_\pi$$
$$= \sup_{\|v\|_\pi = 1} \mathbb{E}_0 \left[\left\langle v, \frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu_0}(X_i) - 1 \right\rangle_\pi \middle| Y \right]$$

Centered likelihood ratio has mean zero,
… but here we condition on Y

- Bound conditional expectation using concentration [Han, Özgür, Weissman, '18]

Related toy problem

- Suppose $Z \sim \mathrm{N}(0,1)$ and $Y = Y(Z)$ is arbitrary function of Z taking at most M possible values
- **Question**: How large can $|\mathbb{E}[Z|Y]|$ (say, in expectation)?
- **Answer**: $O\left(\sqrt{\log M}\right)$
  - For event E, how $|\mathbb{E}[Z|E]|$ depend on $\Pr(E)$?
  - Which event E with $\Pr(E) = \delta$ maximizes $|\mathbb{E}[Z|E]|$?
  - Consider tail event $E = \{Z > \Phi^{-1}(\delta)\}$

In closing…

- In lieu of proving exponential lower bounds for Tensor PCA:
  - We show that current algorithms are unimprovable without worsening some "natural" resource complexity (memory size, sample size, time)
  - Shed light on computational + statistical benefits of overparameterization
  - New communication complexity tools for distributed estimation lower bounds
- Open problems:
  - Algorithms achieving other points on Pareto frontier?
  - Lower bounds for learning problems with higher SNR?

Thank you!

arXiv:2204.07526

Algorithm for Tensor PCA (k=4)

Partial trace algorithm [HSSS'16]

- Let $A \in \mathbb{R}^{d\times d}$ be matrix given by
$$A_{i,j} = \sum_{l=1}^{d} \bar{X}_{i,j,l,l}$$
- Return $\hat{\theta}$ = top eigenvector of A

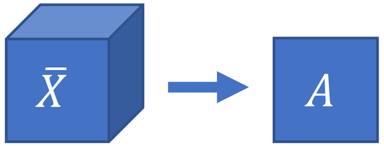
Recall:

$$\bar{X} \sim \lambda\, \theta^{\otimes 4} + \frac{1}{\sqrt{N}} Z$$

Partial trace matrix:

$$A \sim \lambda\, \theta^{\otimes 2} + \frac{1}{\sqrt{N}} \sqrt{d}\, Z'$$

SNR reduced from λ^2 to λ^2/d



$\mathrm{TPCA}(d, 4, \lambda^2) \to \mathrm{TPCA}(d, 2, \lambda^2/d)$

Sample size requirement

$$N \asymp \frac{d}{\lambda^2/d} = \frac{d^2}{\lambda^2}$$

Power method impl.

Memory size: $B \asymp d$

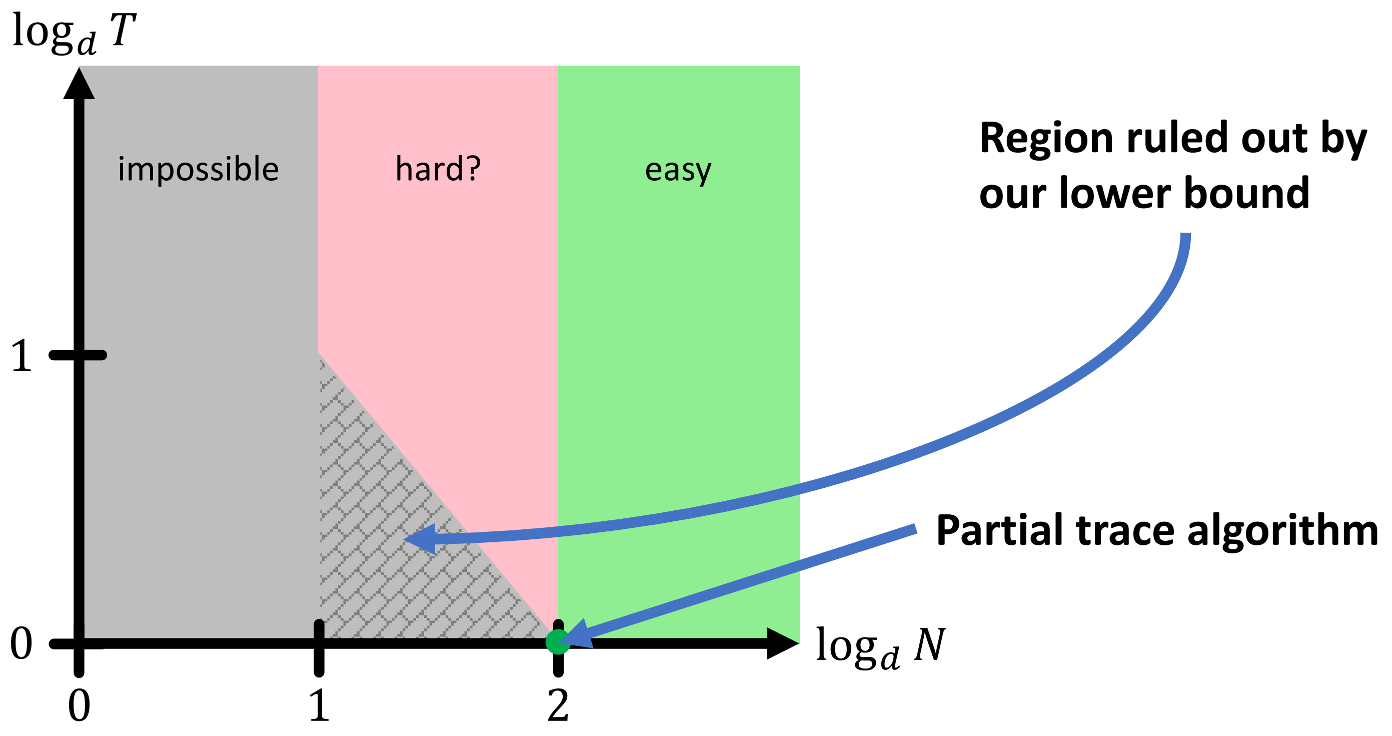
Iterations: $T \asymp \log d$

Total resources

$$BNT \asymp \frac{d^3}{\lambda^2} \log d$$

Phase diagram for Tensor PCA

- Linear memory algorithms with $B \asymp d$



Frameworks for communication lower bounds

- Prior works study "hide-and-seek" variant of estimation problem
  [Shamir, '14; Han, Özgür, Weissman, '18; Acharya, Canonne, Sun, Tyagi, '22]
  - Nature chooses $\theta \sim \pi$ and $J \in [d]$ uniformly at random
  - Data is drawn from μ_θ and distributed to the parties
  - Parameter θ is revealed to all parties except with J-th component re-randomized (and J is kept hidden)
- Hide-and-seek problem is solved with $O(Nd)$ communication
  - Each party sends likelihoods of all $O(d)$ possibilities given own datum
  - Cannot use this to prove lower bounds of $d^{\lceil (k+1)/2 \rceil}$ bits (except if $k = 2$)

Using structure of blackboard protocols

- Leverage special structure of blackboard protocols [Bar-Yossef et al, '04]
  - In $P_\theta^{(i)}$, get transcript Y using $X_i \sim P_\theta$ and $X_j \sim P_0$ for all $j \neq i$:

$$\mathrm{h}^2(P_\theta; P_0) \lesssim \sum_{i=1}^{N} \mathrm{h}^2\left(P_\theta^{(i)}, P_0\right)$$

  - Moreover:

$$\mathrm{h}^2\left(P_\theta^{(i)}, P_0\right) = \mathbb{E}_0\left[\mathbb{E}_0\left[\frac{\mathrm{d}\mu_\theta}{\mathrm{d}\mu_0}(X_i) - 1 \middle| Y\right]^2\right]$$

Solution to toy problem

- For any $\lambda \in (0,0.5)$,

$$\mathbb{E}[\exp(\lambda Z^2)] = (1-2\lambda)^{-1/2} = O(1)$$

- So, conditional on event $Y = y$,

$$\mathbb{E}[\exp(\lambda Z^2)\,|Y = y] = O(1)/\Pr(Y = y)$$

- By Jensen's inequality and convexity of $t \mapsto \exp(\lambda t^2)$,

$$\exp(\lambda \mathbb{E}[Z|Y = y]^2) \le O(1)/\Pr(Y = y)$$

- Rearrange: $\mathbb{E}[Z|Y = y]^2 \le O(\log(1/\Pr(Y = y)))$

Comparison to [Raz, '16]

- [Raz, '16]: Every algorithm for learning d-bit parity functions requires either $\Omega(d^2)$ bits of memory or $2^{\Omega(d)}$ samples
  - Streaming setup: random example is either stored in memory or gone forever
  - Time = sample size
- **Our setup**:
  - We don't count data set towards memory cost
    - Only charge for additional "working memory"
    - [Kong, '18]: d-bit parities can be learned with $O(d)$ samples, $O(d)$ bits of working memory, and $\mathrm{poly}(d)$ passes through data
  - We allow for multiple passes through data set
  - But we require noise, and cannot imply exponential complexity

Parallel Pancakes

$$Y \sim \frac{1}{2}\mathrm{N}(-\mu, I_d - \mu\mu^\top) + \frac{1}{2}\mathrm{N}(\mu, I_d - \mu\mu^\top)$$

Assume $\lambda^2 \coloneqq \|\mu\|^8 < d^{-10}$

If algorithm computes estimate $\hat{\mu}$ satisfying

$$\mathbb{E}\left[\frac{\langle \hat{\mu}, \mu \rangle}{\|\hat{\mu}\|_2 \|\mu\|_2}\right] \gg \frac{1}{\sqrt{d}}$$

then it must use

$$\text{memory size} \times \text{sample size} \times \text{time} \gtrsim \frac{d^3}{\lambda^2}$$