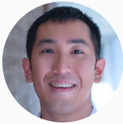
Wedges are all you need: sparser and sparser tensor completion

Ludovic Stephan

ENSAI - CREST

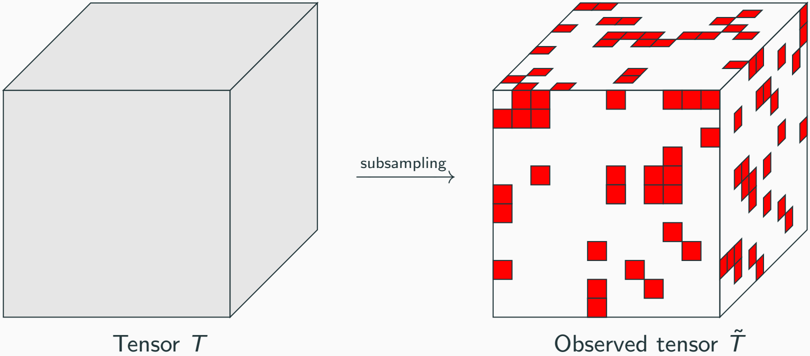


Hengrui Luo
Rice University

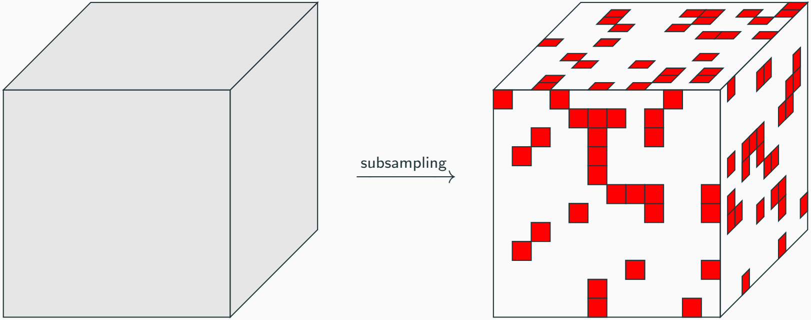
Anna Ma
UC Irvine

Yizhe Zhu
USC

What is tensor completion ?



What is tensor completion ?



Tensor T Observed tensor $\tilde{T}$

- T is an order-k tensor of size $n \times \cdots \times n$
- The observed tensor $\tilde{T}$ is defined as

$$\tilde{T}_{i_1,\ldots,i_k} = \begin{cases} T_{i_1,\ldots,i_k} & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

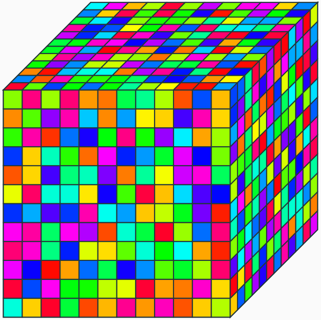
- Goal: Exactly/approximately recover T from $\tilde{T}$ with very few samples (with an efficient algorithm)

Why do we care?

Typical appplications: recommendation systems

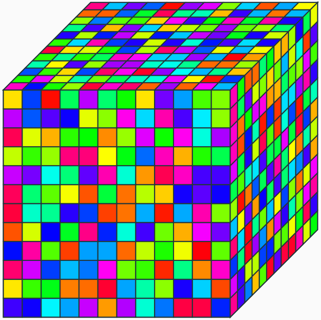
- Each axis represents a *modality*: users, movies/food, time of day...
- Revealed entries are *feedback*, e.g. ratings
- Goal: predict how a (new) user will rate an item at a specific time

Model assumptions

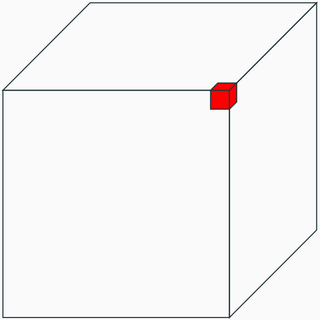


Too many degrees of freedom!

Model assumptions

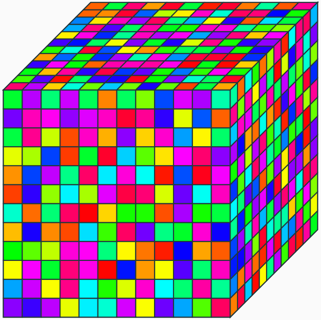


Too many degrees of freedom!

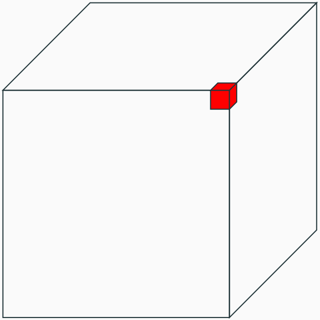


Too localized!

Model assumptions



Too many degrees of freedom!



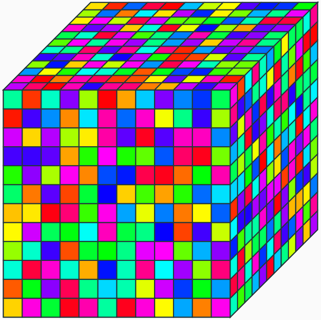
Too localized!

- T has low CP-rank:

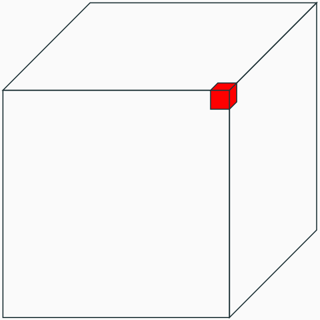
$$T = \sum_{i=1}^{r} \lambda_i \left(w_i^{(1)} \otimes \cdots \otimes w_i^{(k)} \right)$$

$\Rightarrow r \times kn$ degrees of freedom

Model assumptions



Too many degrees of freedom!



Too localized!

- T has low CP-rank:

$$T = \sum_{i=1}^{r} \lambda_i \left(w_i^{(1)} \otimes \cdots \otimes w_i^{(k)} \right)$$

$\Rightarrow r \times kn$ degrees of freedom

- T is delocalized:

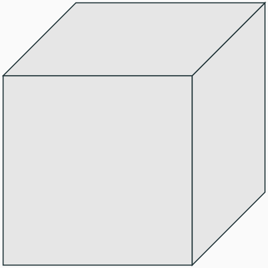
$$\|w_i^{(j)}\|_\infty \simeq n^{-1/2}$$

Computational hardness

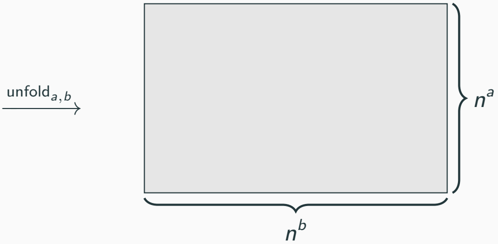
Computational complexity problem: most tensor problems are hard [Hillar-Lim '09]

- spectral norm
- eigenvalues/singular values
- low-rank approximations

Unfolding



k-tensor (size $n \times \cdots \times n$)

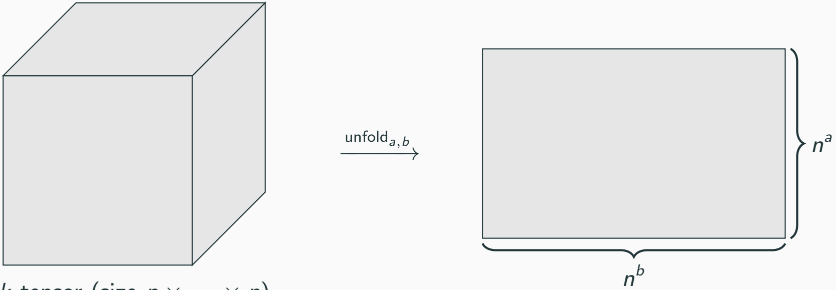


Unfolding matrix (size $n^a \times n^b$)

"Grouping" indices:

$$M_{(i_1,\ldots,i_a),(i_{a+1},\ldots,i_k)} = T_{i_1,\ldots,i_k}$$

Unfolding



k-tensor (size $n \times \cdots \times n$)

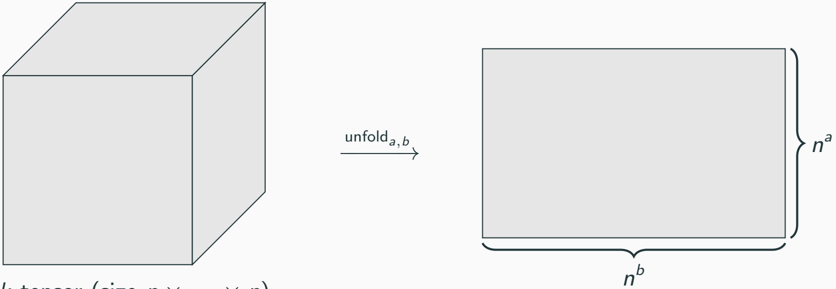
Unfolding matrix (size $n^a \times n^b$)

"Grouping" indices:

$$M_{(i_1,\ldots,i_a),(i_{a+1},\ldots,i_k)} = T_{i_1,\ldots,i_k}$$

Tensor completion on $T \Leftarrow$ Matrix completion on M

Unfolding



k-tensor (size $n \times \cdots \times n$)

Unfolding matrix (size $n^a \times n^b$)

"Grouping" indices:

$$M_{(i_1,\ldots,i_a),(i_{a+1},\ldots,i_k)} = T_{i_1,\ldots,i_k}$$

Tensor completion on $T \Leftarrow$ Matrix completion on M

If k is even: square matrix of size $n^{k/2} \implies \tilde{O}(n^{k/2})$ samples suffice
If k is odd: matrix of size $n^{\lfloor k/2 \rfloor} \times n^{\lceil k/2 \rceil}$

Statistical-computational gap for random tensors

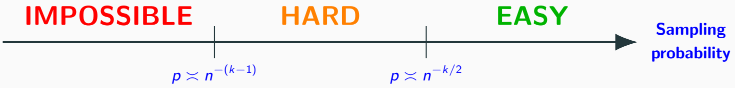
- NP-hard algorithms: tensor-based norm minimization methods without unfolding
  [Yuan and Zhang '16, Ghadermarzy et al '19, Harris and Zhu '21]
  $\rightarrow$ works with $\tilde{O}(n)$ samples

Statistical-computational gap for random tensors

- NP-hard algorithms: tensor-based norm minimization methods without unfolding
  [Yuan and Zhang '16, Ghadermarzy et al '19, Harris and Zhu '21]
  $\rightarrow$ works with $\tilde{O}(n)$ samples
- Unfolding-based algorithms with spectral initialization
  [Montanari and Sun '16, Liu and Moitra '20, Cai et al. '21...]
  $\rightarrow$ works with $\tilde{O}(n^{k/2})$ samples

Statistical-computational gap for random tensors

- NP-hard algorithms: tensor-based norm minimization methods without unfolding
  [Yuan and Zhang '16, Ghadermarzy et al '19, Harris and Zhu '21]
  $\rightarrow$ works with $\tilde{O}(n)$ samples
- Unfolding-based algorithms with spectral initialization
  [Montanari and Sun '16, Liu and Moitra '20, Cai et al. '21...]
  $\rightarrow$ works with $\tilde{O}(n^{k/2})$ samples
- Similar gaps in the spiked tensor model $T = \lambda v^{\otimes q} + Z$
  [Montanari and Richard '14, Ben Arous et al. '17, Chen '18, Ben Arous et al. '18, Wein et al. '19, Perry et al. '20...]



Basic unfolding algorithm

Commonly poly-time algorithms: unfolding-based

- Unfold $\tilde{T}$ into $A \in \mathbb{R}^{n \times n^2}$
- Spectral initialization: truncated SVD of the hollowed matrix $AA^\top - \mathrm{diag}(AA^T)$
- Post-processing: projection [Montanari and Sun '18], tensor power iteration [Xia et al '21], gradient descent [Xia and Yuan '19, Cai et al. '21]

$\rightarrow$ works until $p = O(n^{-k/2} \times \text{polylog}(n))$

Basic unfolding algorithm

Commonly poly-time algorithms: unfolding-based

- Unfold $\tilde{T}$ into $A \in \mathbb{R}^{n \times n^2}$
- Spectral initialization: truncated SVD of the hollowed matrix $AA^\top - \mathrm{diag}(AA^T)$
- Post-processing: projection [Montanari and Sun '18], tensor power iteration [Xia et al '21], gradient descent [Xia and Yuan '19, Cai et al. '21]

$\rightarrow$ works until $p = O(n^{-k/2} \times \mathrm{polylog}(n))$

What happens if $p \propto n^{-k/2}$?

Not a trivial challenge

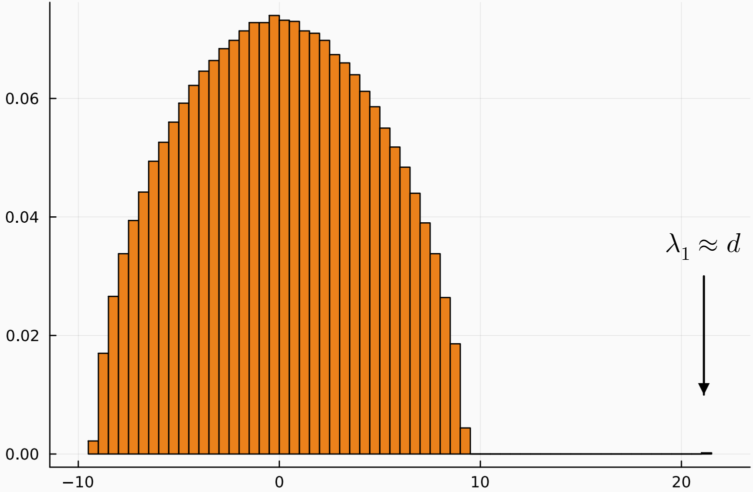


Figure: $T = v \otimes v \otimes v, AA^\top - \mathrm{diag}(AA^\top), p = 20n^{-3/2}$

Not a trivial challenge

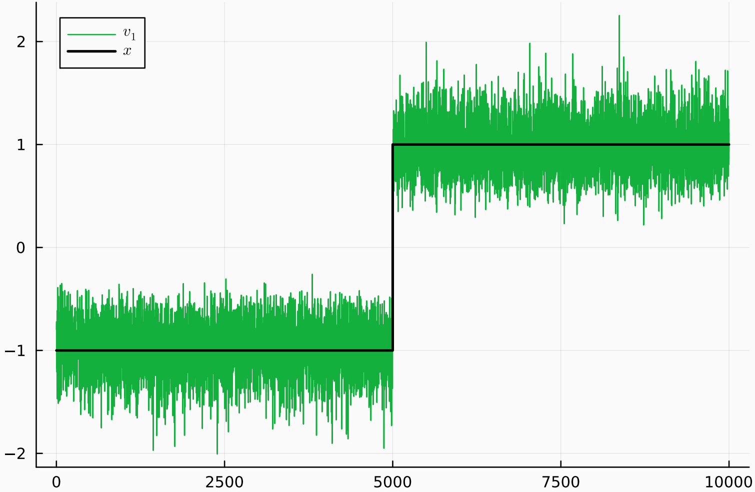


Figure: $T = v \otimes v \otimes v, AA^\top - \mathrm{diag}(AA^\top), p = 20n^{-3/2}$

Not a trivial challenge

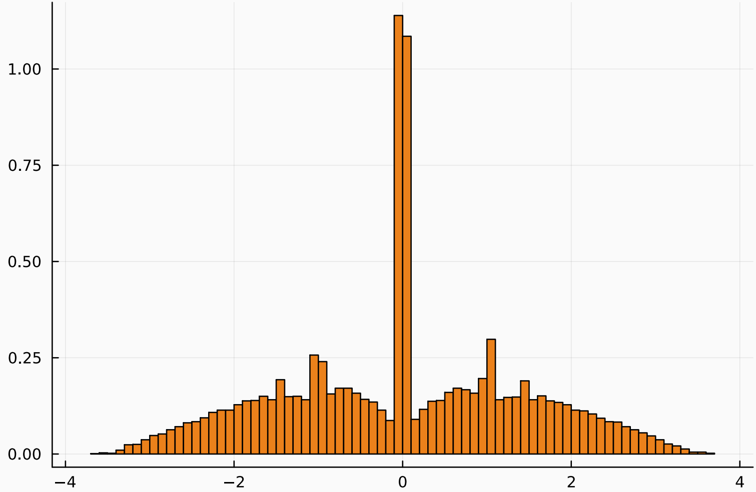


Figure: $AA^\top - \mathrm{diag}(AA^\top), p = 2n^{-3/2}$

Not a trivial challenge

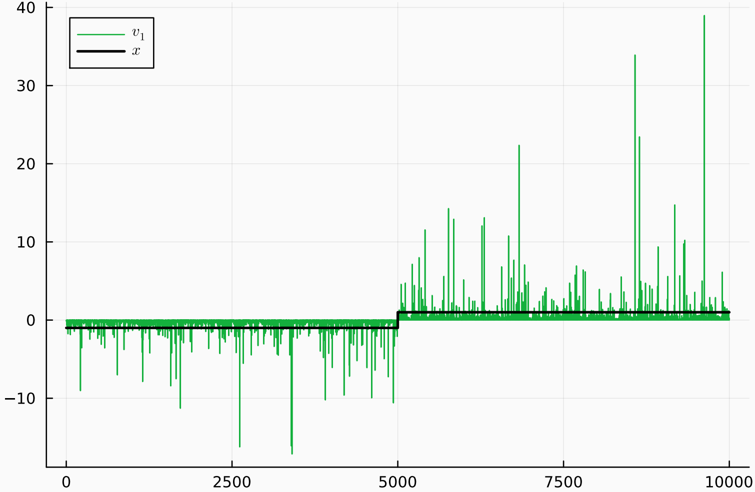
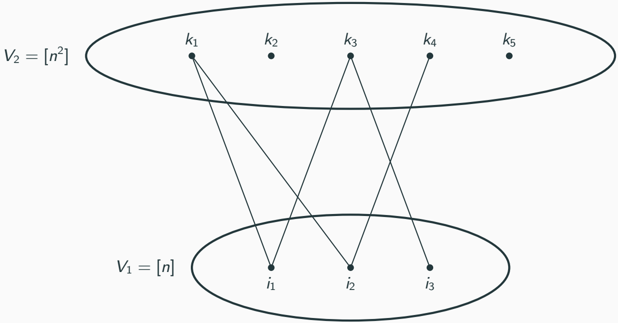


Figure: $AA^\top - \mathrm{diag}(AA^\top), p = 2n^{-3/2}$

A random graph theory explanation

$A \in \mathbb{R}^{n \times n^2}$ corresponds to a (weighted) random bipartite graph with $V_1 = [n]$, $V_2 = [n^2]$.

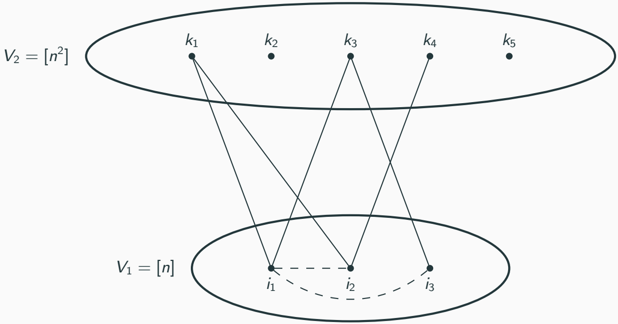


A random graph theory explanation

Hollowed matrix counts walks of length 2, $V_1 \to V_2 \to V_1$:

$$(AA^\top)_{ij} = \sum_k A_{ik} A_{jk}.$$

$h(AA^\top)$ can be seen as the adjacency matrix of a new graph $\tilde{G}$ (dashed edges).



A random graph theory explanation

Fact: $\tilde{G}$ is still sparse (average degree d^2 for $p = dn^{-k/2}$).

In the unweighted (Erdős-Rényi) case:

- if $d^2 \gtrsim \sqrt{\frac{\log(n)}{\log\log(n)}}$: spectrum of $\tilde{G}$ concentrates [Feige and Ofek '05, Benaych-Georges et al. '20]
- if $d^2 \ll \sqrt{\frac{\log(n)}{\log\log(n)}}$: no concentration, spectrum dominated by high-degree vertices [Benaych-Georges et al. '19]

A random graph theory explanation

Fact: $\tilde{G}$ is still sparse (average degree d^2 for $p = dn^{-k/2}$).

In the unweighted (Erdős-Rényi) case:

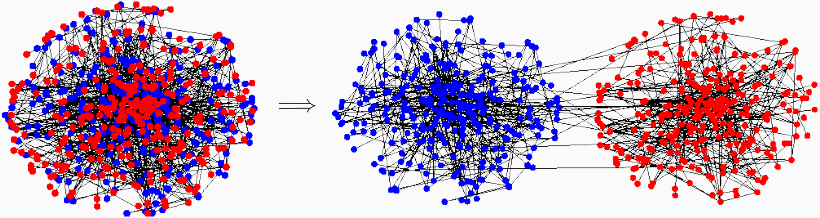
- if $d^2 \gtrsim \sqrt{\frac{\log(n)}{\log\log(n)}}$: spectrum of $\tilde{G}$ concentrates [Feige and Ofek '05, Benaych-Georges et al. '20]
- if $d^2 \ll \sqrt{\frac{\log(n)}{\log\log(n)}}$: no concentration, spectrum dominated by high-degree vertices [Benaych-Georges et al. '19]

$\Rightarrow$ Naive unfolding (probably) doesn't work

A detour through community detection

Community detection in stochastic block models $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$.

- Unknown partition $\sigma \in \{-1, 1\}^n$. Generate a random graph $G = ([n], E)$. i, j is connected with probability $p = \frac{a}{n}$ if $\sigma_i = \sigma_j$ and with probability $q = \frac{b}{n}$ otherwise.
- goal: recover σ from G

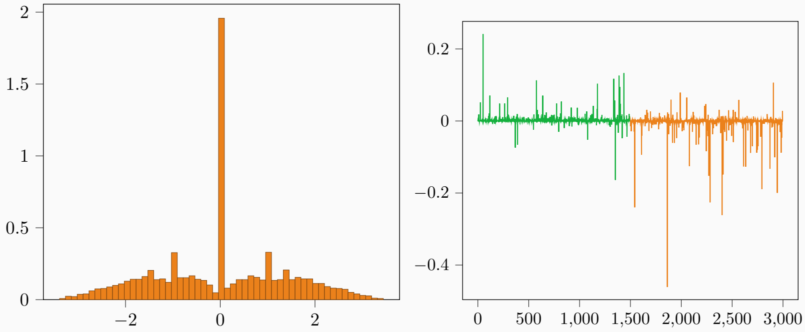


A detour through community detection

$\mathbb{E}[A]$ is low-rank, and $v_2(\mathbb{E}[A]) = \sigma \Rightarrow$ spectral method on A?

A detour through community detection

$\mathbb{E}[A]$ is low-rank, and $v_2(\mathbb{E}[A]) = \sigma \Rightarrow$ spectral method on A? **No!**



High-degree vertices dominate the spectrum. v_2 localized around high-degree vertices.

[Krivelevich and Sudakov '01, Benaych-Georges et al. '19, Alt et al. '23]

Non-backtracking matrix for graphs

Proposed in [Krzakala et al. '13]

Defined on the oriented edges of G:

$$\vec{E} = \{u \to v : \{u, v\} \in E\}, |\vec{E}| = 2|E|.$$

Non-backtracking matrix for graphs

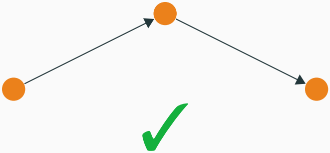
Proposed in [Krzakala et al. '13]

Defined on the oriented edges of G:

$$\vec{E} = \{u \to v : \{u, v\} \in E\}, |\vec{E}| = 2|E|.$$

The non-backtracking matrix B is defined: for $u \to v, x \to y \in \vec{E}$,

$$B_{u \to v, x \to y} = \mathbf{1}_{v=x} \mathbf{1}_{u \neq y}.$$



Non-backtracking matrix for graphs

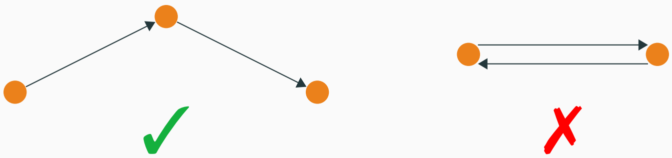
Proposed in [Krzakala et al. '13]

Defined on the oriented edges of G:

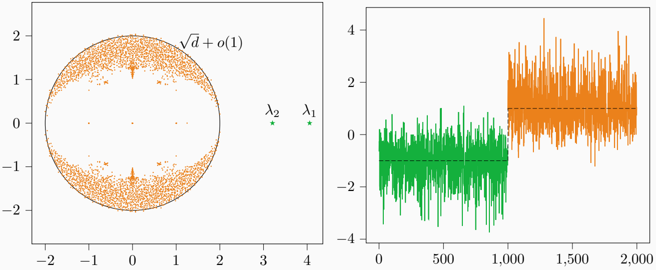
$$\vec{E} = \{u \to v : \{u, v\} \in E\}, |\vec{E}| = 2|E|.$$

The non-backtracking matrix B is defined: for $u \to v, x \to y \in \vec{E}$,

$$B_{u\to v, x\to y} = \mathbf{1}_{v=x}\mathbf{1}_{u\neq y}.$$

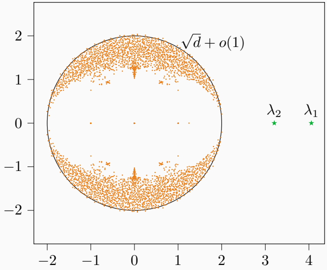


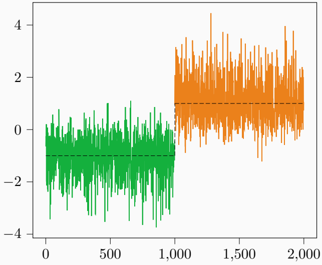
Non-backtracking spectral method



- If $(a-b)^2 > 2(a+b)$, then the second eigenvector of B can be used to detect the community structure. [Bordenave, Lelarge, Massoulié '18]

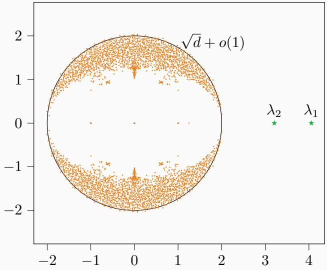
Non-backtracking spectral method

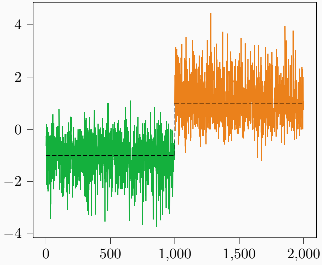




- If $(a-b)^2 > 2(a+b)$, then the second eigenvector of B can be used to detect the community structure. [Bordenave, Lelarge, Massoulié '18]
- B is non-Hermitian: avoid the localization effect from high degree vertices when G is very sparse.

Non-backtracking spectral method





- If $(a-b)^2 > 2(a+b)$, then the second eigenvector of B can be used to detect the community structure. [Bordenave, Lelarge, Massoulié '18]
- B is non-Hermitian: avoid the localization effect from high degree vertices when G is very sparse.
- Can be generalized for *estimating a low-rank structure from sparse observations* with $O(n)$ many samples [S.-Massoulié '23].
  In particular: very sparse matrix completion !
  [Bordenave-Coste-Nadakuditi '23]

A new non-backtracking matrix for sparse long matrices

Long matrix completion

- Rectangular matrix M of size $n \times m$ ($m \gg n$), with SVD

$$M = \sum_{i=1}^{r} \nu_i \phi_i \psi_i^\top, \quad MM^\top = \sum_{i=1}^{r} \nu_i^2 \phi_i \phi_i^\top$$

- Masking matrix X with $X_{ij} \sim \mathrm{Ber}(p)$, $p = \frac{d}{\sqrt{mn}}$
- Observed matrix:

$$A = \frac{X \circ M}{p} \quad \text{so that} \quad \mathbb{E}[A] = M$$

Long matrix completion

- Rectangular matrix M of size $n \times m$ ($m \gg n$), with SVD

$$M = \sum_{i=1}^{r} \nu_i \phi_i \psi_i^\top, \quad MM^\top = \sum_{i=1}^{r} \nu_i^2 \phi_i \phi_i^\top$$

- Masking matrix X with $X_{ij} \sim \mathrm{Ber}(p)$, $p = \frac{d}{\sqrt{mn}}$
- Observed matrix:

$$A = \frac{X \circ M}{p} \quad \text{so that} \quad \mathbb{E}[A] = M$$

Assumptions:

$$r, \sqrt{n}\|\phi_i\|_\infty = O(\mathrm{polylog}(n))$$

Goal: estimate singular values and left singular vectors of M: ν_i, ϕ_i, with sample size $O(\sqrt{mn})$

Long matrix completion

- Rectangular matrix M of size $n \times m$ ($m \gg n$), with SVD

$$M = \sum_{i=1}^{r} \nu_i \phi_i \psi_i^\top, \quad MM^\top = \sum_{i=1}^{r} \nu_i^2 \phi_i \phi_i^\top$$

- Masking matrix X with $X_{ij} \sim \mathrm{Ber}(p)$, $p = \frac{d}{\sqrt{mn}}$
- Observed matrix:

$$A = \frac{X \circ M}{p} \quad \text{so that} \quad \mathbb{E}[A] = M$$

Assumptions:

$$r, \sqrt{n}\|\phi_i\|_\infty = O(\mathrm{polylog}(n))$$

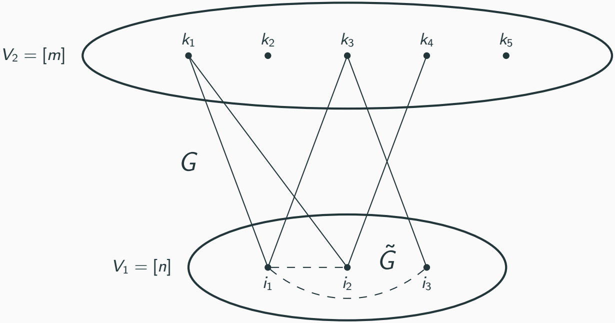
Goal: estimate singular values and left singular vectors of M: ν_i, ϕ_i, with sample size $O(\sqrt{mn})$

Estimating the full SVD of M needs $\Omega(m)$ samples [Koltchinskii et al. '11]

Graph and folded graph

G: bipartite graph on $V_1 \times V_2$, adjacency matrix X

$\tilde{G}$: (multi)-graph on V_1, adjacency matrix $XX^\top$



Non-backtracking wedge matrix

First idea: take the (weighted) non-backtracking matrix of $\tilde{G}$ $\Rightarrow$ doesn't work

Non-backtracking wedge matrix

First idea: take the (weighted) non-backtracking matrix of $\tilde{G}$ $\Rightarrow$ doesn't work

Better idea: work directly on *oriented wedges* in G

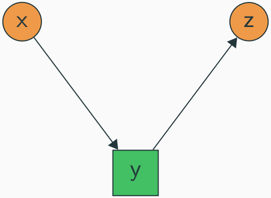
$$\vec{E}_2 = \{(x, y, z) \in V_1 \times V_2 \times V_1, z \neq x\}$$

Non-backtracking wedge matrix

First idea: take the (weighted) non-backtracking matrix of $\tilde{G} \Rightarrow$ doesn't work

Better idea: work directly on *oriented wedges* in G

$$\vec{E}_2 = \{(x, y, z) \in V_1 \times V_2 \times V_1, z \neq x\}$$

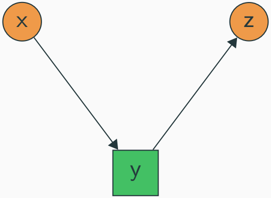


Non-backtracking wedge matrix

First idea: take the (weighted) non-backtracking matrix of $\tilde{G}$ $\Rightarrow$ doesn't work

Better idea: work directly on *oriented wedges* in G

$$\vec{E}_2 = \{(x, y, z) \in V_1 \times V_2 \times V_1, z \neq x\}$$



$\Rightarrow$ $\vec{E}_2$ has size $\sim n^2 m p^2 = d^2 n$: independent from m

Non-backtracking wedge matrix

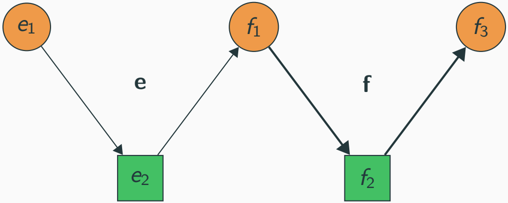
Define B indexed by $\vec{E}$ as

$$B_{ef} = \begin{cases} A_{f_1 f_2} A_{f_3 f_2} & \text{if } e_3 = f_1 \text{ and } e_2 \neq f_2 \\ 0 & \text{otherwise} \end{cases}$$

Non-backtracking wedge matrix

Define B indexed by $\vec{E}$ as

$$B_{ef} = \begin{cases} A_{f_1 f_2} A_{f_3 f_2} & \text{if } e_3 = f_1 \text{ and } e_2 \neq f_2 \\ 0 & \text{otherwise} \end{cases}$$



e, f form a non-backtracking walk of length 4, starting from V_1, ending in V_1.

Results

Theorem (S.-Zhu '24)

Assume that $p \propto \frac{d}{\sqrt{mn}}$*, with* d *large enough. Then with high probability, the top eigenvalues (resp. eigenvectors) of* B *are correlated with the* ν_i *and* ϕ_i*, in the sense that we can build estimates* λ_i*,* ξ_i *satisfying*

$$\lambda_i = \nu_i + o(1) \quad \text{and} \quad \langle \xi_i, \phi_i \rangle^2 = 1 - O\left(\frac{1}{d}\right) + o(1)$$

When $M = \mathrm{unfold}(T)$, we can achieve *weak recovery* of T, and *almost exact* recovery ($\|T - \hat{T}\| = o(1)$) when $d \to \infty$!

Results: eigenvalues

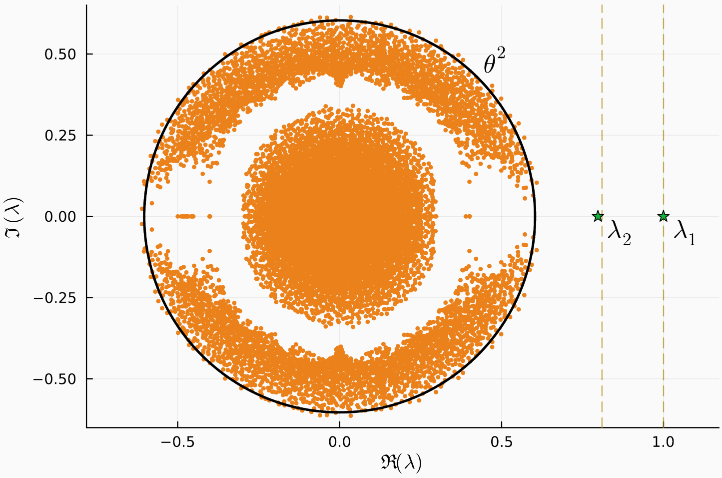


Figure: Spectrum of $B, d = 3$

Results: eigenvectors

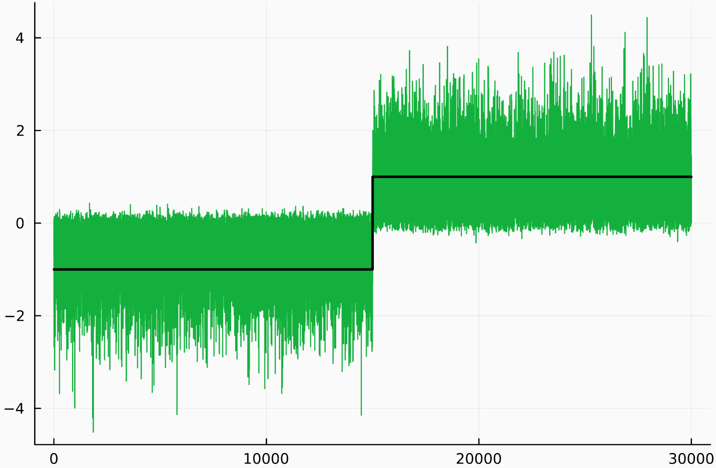


Figure: Top eigenvector of $B, d = 3$

Wedge sampling

Lessons from before

Studying $AA^\top$ $\Leftrightarrow$ studying wedges

All non-wedges (vertices in V_2 with degree 1) are useless!

Uniform sampling: $O(n^{k/2})$ degree-one vertices, $O(n)$ wedges...

Lessons from before

Studying $AA^\top$ $\Leftrightarrow$ studying wedges

All non-wedges (vertices in V_2 with degree 1) are useless!

Uniform sampling: $O(n^{k/2})$ degree-one vertices, $O(n)$ wedges...

What if we could only sample 'useful' edges?

Wedge sampling

Sample from the *wedge set*

$$\mathcal{W} = [n] \times [m] \times [n]$$

with probability $p = \frac{1}{mn} \times \text{polylog}(n) \to$ sampled set $\tilde{\mathcal{W}}$

For each wedge (i, k, j), reveal A_{ik} and A_{jk}

mn^2 possible wedges, $p \lesssim \frac{1}{mn}$, two entries revealed per wedge
$\Rightarrow$ $\tilde{O}(n)$ samples!

Spectral initialization

Initial spectral estimator:

$$\tilde{B} = \sum_{(i,k,j)\in\tilde{\mathcal{W}}} A_{ik}A_{jk}.$$

Theorem (Luo, Ma, S., Zhu '25)

Let $p \gtrsim \frac{1}{mn}$ and $A \in \mathbb{R}^{n\times m}$ a low-rank delocalized matrix. Then with high probability,

$$\|\tilde{B} - AA^\top\| \lesssim \sqrt{\frac{\log(n)}{pnm}}\|A\|^2$$

Further, if $A = U\Sigma V^\top$ and $\tilde{B} = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$, then

$$\min_{O\in\mathcal{O}_r}\|\tilde{U}_r O - U\|_{2,\infty} \lesssim \sqrt{\frac{\log(n)}{pnm}}\|U\|_{2,\infty}$$

where $\tilde{U}_r$ contains the top r eigenvectors of $\tilde{B}$.

Ilustration

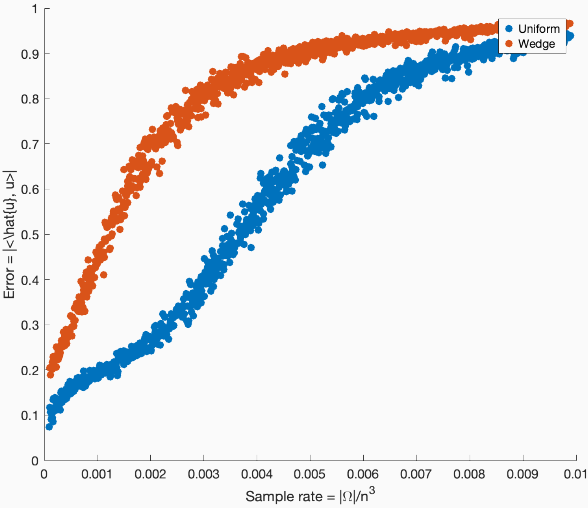


Figure: Recovery performance for a rank-one tensor

Refinement

Previous works [Haselby et al. '24]: end-to-end algorithm adapted to the sampling scheme.

Our idea: sample splitting

- wedge sampling for initialization
- sparse uniform sampling ($p \propto n^{-(k-1)}$) with existing algorithms [Montanari and Sun '18, Cai et al. '21]

Theorem (Luo, Ma, S., Zhu '25)

Once the initial estimate $\tilde{U}_r$ of U is obtained, the refinement methods of [Montanari and Sun '18, Cai et al '21] only require $\tilde{O}(n)$ uniform samples to recover T

Hardest step is initial alignment [Ben Arous et al '21]

Under the hood

Core property for the refinement step: concentration in spectral norm

$$\|p^{-1}\tilde{T} - T\| \lesssim \frac{1}{n^{k/2}p}\|T\|$$

Problem: this is sharp!

$$\|p^{-1}\tilde{T}\| \gtrsim \|p^{-1}\tilde{T}\|_\infty \gtrsim \frac{1}{n^{k/2}p}\|T\|$$

$\Rightarrow$ no concentration for $p \ll n^{-k/2}$...

Delocalized norm of tensors

Main idea:

$$\|T\| = \max_{\|u_i\| \leq 1} \langle T, u_1 \otimes \cdots \otimes u_k \rangle,$$

but

- $\|p^{-1}\tilde{T}\|$ is attained for $u_i = e_{j_i}$ (basis vectors)
- in the proofs, usually *almost all* of the u_i are delocalized!

New norm: for $\delta \in \mathbb{R}^k$,

$$\|T\|_\delta = \sup_{j_1, j_2 \in [k]} \sup_{(u_1, \ldots, u_k) \in \mathcal{U}_{j_1 j_2}} \langle T, u_1 \otimes \cdots \otimes u_k \rangle$$

where

$$\mathcal{U}_{j_1 j_2} = \{(u_1, \ldots, u_k) : \|u_i\| \leq 1 \ \forall i, \|u_j\|_\infty \leq \delta_j \ \forall j \neq j_1, j_2\}$$

Delocalized norm concentration

Concentration on much sparser tensors for $\|\cdot\|_\delta$:

Theorem (Yuan and Zhang '17, Luo, Ma, S., Zhu '25)

Assume that $\delta_i \lesssim n^{-1/2}$ for all $i \in [k]$. Then for any low-rank delocalized tensor T, with high probability

$$\|p^{-1}\tilde{T} - T\|_\delta \lesssim \sqrt{\frac{\log(d)}{n^{k-1}p}}\|T\|_\delta \lesssim \sqrt{\frac{\log(d)}{n^{k-1}p}}\|T\|$$

Already used for non-polynomial tensor completion, never for the polynomial case!

Is it viable?

Not only viable... but *natural*!

Sampling $(i, \underline{k}, j) \Leftrightarrow$ fixing all modalities but one

Core principle of experimental design!

Thank you!