Computation-information trade-offs in algorithms for Low-Rank Tensor Sensing

Igor Molybog

Plan

1. Tensor PCA
   - SOS hierarchy
   - Kikuchi Hierarchy
   - Local search + Restarts

Plan

1. Tensor PCA
   - SOS hierarchy
   - Kikuchi Hierarchy
   - Local search + Restarts
2. Matrix sensing
   - Matrix sensing under RIP
   - Matrix completion
   - Adding noise

Rank-r p-tensor recovery with noise

Measurement :

$$Y = \lambda \, \mathcal{A} \left(\sum_{i=1}^{r} u_i^{\otimes p} \right) + W$$

where $\mathcal{A}$ is a linear operator, W is a noise term (standard Gaussian) ;
λ, r, p — parameters

Rank-r p-tensor recovery with noise

Measurement :

$$Y = \lambda \mathcal{A} \left(\sum_{i=1}^{r} u_i^{\otimes p} \right) + W$$

where $\mathcal{A}$ is a linear operator, W is a noise term (standard Gaussian) ;
λ, r, p — parameters

Applications :

- Recommender systems
- Power grid analysis
- Quantum state preparation and tomography
- Logistics
- Neural architecture search

Tensor PCA MLE problem formulation

- Trivial sensing operator : $\mathcal{A} = I$
- Rank $r = 1$

$$Y = \lambda\, u^{\otimes p} + W$$

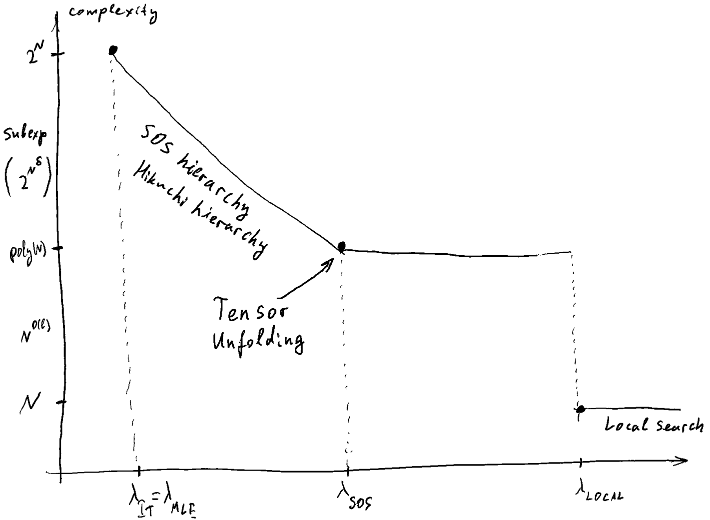
Likelyhood under Gaussian noise :

$$f(x) = \lambda \langle u, x \rangle^p - H_{n,p}(x)$$

Domains :

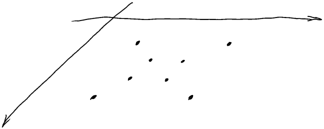
- $x, u \in \mathbb{S}^{n-1}$
- $x, u \in \{\pm 1\}^n$

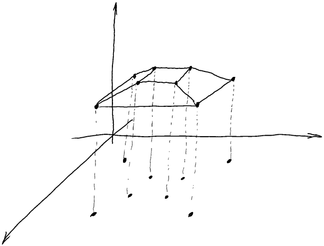
Complexity of Tensor PCA



Idea of lifting

- The extreme values of a concave function are reached at extreme points of the domain
- $\pi(\mathit{ext}[\mathit{Lifted}]) = \mathit{ext}[\mathit{Unlifted}]$





Sum of Squares

$$P \geq 0 \Leftrightarrow P = \sum_i \left(\frac{P_i}{Q_i}\right)^2$$

Sum of Squares

$$P \geq 0 \Leftrightarrow P = \sum_i \left(\frac{P_i}{Q_i}\right)^2$$

Theorem (Corollary of Positivstellensatz (Krivine '64 and Stengle '74))

Given $P_1, \ldots, P_m \in \mathbb{R}[x_1, \ldots, x_n] = \mathbb{R}[x]$

$$\{P_1 = 0, \ldots P_m = 0\} = \varnothing \quad \iff \quad -1 \equiv S + \sum_i Q_i P_i$$

for some certificate

- $S, Q_1, \ldots, Q_m \in \mathbb{R}[x]$
- *S is a sum of squares of polynomials*

Sum of Squares

$$P \geq 0 \Leftrightarrow P = \sum_i \left(\frac{P_i}{Q_i}\right)^2$$

Theorem (Corollary of Positivstellensatz (Krivine '64 and Stengle '74))

Given $P_1, \ldots, P_m \in \mathbb{R}[x_1, \ldots, x_n] = \mathbb{R}[x]$

$$\{P_1 = 0, \ldots P_m = 0\} = \varnothing \quad \Longleftrightarrow \quad -1 \equiv S + \sum_i Q_i P_i$$

for some certificate

- $S, Q_1, \ldots, Q_m \in \mathbb{R}[x]$
- *S is a sum of squares of polynomials*

- The degree of the certificate ℓ is the maximal degree of $P_i Q_i$

Sum of Squares

$$P \geq 0 \Leftrightarrow P = \sum_i \left(\frac{P_i}{Q_i}\right)^2$$

Theorem (Corollary of Positivstellensatz (Krivine '64 and Stengle '74))

Given $P_1, \ldots, P_m \in \mathbb{R}[x_1, \ldots, x_n] = \mathbb{R}[x]$

$$\{P_1 = 0, \ldots P_m = 0\} = \varnothing \qquad \Longleftrightarrow \qquad -1 \equiv S + \sum_i Q_i P_i$$

for some certificate

- $S, Q_1, \ldots, Q_m \in \mathbb{R}[x]$
- *S is a sum of squares of polynomials*

- The degree of the certificate ℓ is the maximal degree of $P_i Q_i$
- Parametrization of a polynomial of $\deg(S) = \ell$:
  $S(x) = \sum_{\alpha,\alpha'} M_{\alpha\alpha'} x^\alpha x^{\alpha'}$, where $|\alpha|, |\alpha'| \leq \ell/2$

Sum of Squares

$$P \geq 0 \Leftrightarrow P = \sum_i \left(\frac{P_i}{Q_i}\right)^2$$

Theorem (Corollary of Positivstellensatz (Krivine ’64 and Stengle ’74))

Given $P_1, \ldots, P_m \in \mathbb{R}[x_1, \ldots, x_n] = \mathbb{R}[x]$

$$\{P_1 = 0, \ldots P_m = 0\} = \varnothing \qquad \Longleftrightarrow \qquad -1 \equiv S + \sum_i Q_i P_i$$

for some certificate

- $S, Q_1, \ldots, Q_m \in \mathbb{R}[x]$
- *S is a sum of squares of polynomials*

- The degree of the certificate ℓ is the maximal degree of $P_i Q_i$
- Parametrization of a polynomial of $\deg(S) = \ell$:
  $S(x) = \sum_{\alpha,\alpha'} M_{\alpha\alpha'} x^\alpha x^{\alpha'}$, where $|\alpha|, |\alpha'| \leq \ell/2$
- S is a sum of squares if and only if M is PSD ($M \succeq 0$)

Sum of Squares hierarchy

Theorem (Shor '87, Nesterov '00, Parrilo '00, Lasserre '01)

If there exists a degree-ℓ certificate of $\{P_i = 0\}_{i=1}^{m} = \varnothing$ then it can be found in $mn^{O(\ell)}$ time (trough SDP).

Sum of Squares hierarchy

Theorem (Shor '87, Nesterov '00, Parrilo '00, Lasserre '01)

If there exists a degree-ℓ certificate of $\{P_i = 0\}_{i=1}^m = \varnothing$ then it can be found in $mn^{O(\ell)}$ time (trough SDP).

$$\min_{x \in \mathbb{R}^n, P_1(x) = \ldots = P_m(x) = 0} P_0(x) \quad \Leftrightarrow \quad \max \left\{ \psi \mid \{P_0 + \psi = 0, P_i = 0\} = \varnothing \right\}$$

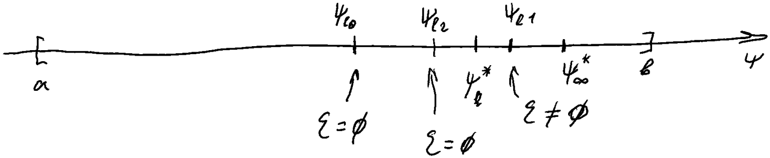
Algorithm :

1. Select level $\ell \in \{0, 1, \ldots\}$
2. Using bisection for ψ over a large interval, ℓ-certify

$$\{P_0 + \psi = 0, P_i = 0\} = \varnothing$$

Sum of Squares hierarchy

$$\mathcal{E}(\psi) = \{P_0 + \psi = 0,\ P_i = 0\} \quad ; \quad \psi^* = \max\{\psi \mid \mathcal{E}(\psi) = \emptyset\}$$



Duality in Sum of Squares

- $\mathbb{R}[x]_\ell = \{P \in \mathbb{R}[x] \mid \deg(P) \le \ell\}$

Duality in Sum of Squares

- $\mathbb{R}[x]_\ell = \{P \in \mathbb{R}[x] \mid \deg(P) \leq \ell\}$
- Lin operator $\bar{\mathbb{E}} : \mathbb{R}[x]_\ell \to \mathbb{R}$ is a degree-ℓ pseudoexpectation if
  - $\bar{\mathbb{E}}[1] = 1$ and
  - $\bar{\mathbb{E}}[P^2] \geq 0$ for all $P \in \mathbb{R}[x]_{\frac{\ell}{2}}$

A pseudoexpectation is representable with a PSD matrix subject to a linear constraint.

Duality in Sum of Squares

- $\mathbb{R}[x]_\ell = \{P \in \mathbb{R}[x] \mid \deg(P) \leq \ell\}$
- Lin operator $\bar{\mathbb{E}} : \mathbb{R}[x]_\ell \to \mathbb{R}$ is a degree-ℓ pseudoexpectation if
  - $\bar{\mathbb{E}}[1] = 1$ and
  - $\bar{\mathbb{E}}[P^2] \geq 0$ for all $P \in \mathbb{R}[x]_{\frac{\ell}{2}}$

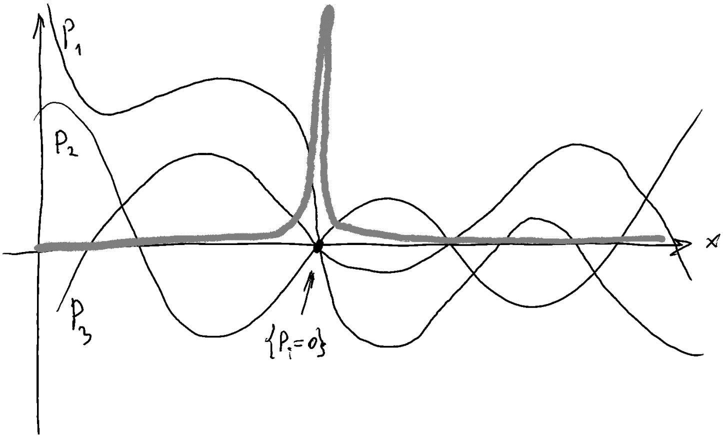
A pseudoexpectation is representable with a PSD matrix subject to a linear constraint.

Theorem (Of SoS Alternatives)

$\{P_i = 0\}_{i=1}^m$ is explicitly bounded. Exactly one holds :

- *Exist a degree-ℓ certificate of $\{P_i = 0\}_{i=1}^m = \varnothing$*
- *Exist a degree-ℓ pseudoexpectation $\bar{\mathbb{E}}$ such that $\bar{\mathbb{E}}(QP_i) = 0$ for all i and Q with $\deg(QP_i) \leq \ell$*

SoS Altrnatives intuition



SoS for Tensor PCA

Minimizing $f \in \mathbb{R}[x]$ subject to $\|x\|^2 - 1 = 0$.

SoS for Tensor PCA

Minimizing $f \in \mathbb{R}[x]$ subject to $\|x\|^2 - 1 = 0$.
Algorithm [Hopkins, Shi, Steurer '15] :

$$\max_{\bar{\mathbb{E}} \in \mathcal{E}} \bar{\mathbb{E}} f$$

where $\mathcal{E} = \left\{ \text{degree-}\ell \;\; \bar{\mathbb{E}} \;\middle|\; \bar{\mathbb{E}}[Q \cdot \{\|x\|^2 - 1\}] = 0 \text{ for all } Q \in \mathbb{R}[x]_{\ell-2} \right\}$
Output :

$$\frac{\bar{\mathbb{E}}^* x}{\|\bar{\mathbb{E}}^* x\|}$$

Complexity result

Theorem (Bhattiprolu, Guruswami, Lee)

For any $1 \leq \ell \leq n$ *if*

$$\lambda \geq \left(\frac{n}{\ell}\right)^{\frac{p-2}{4}} \textit{polylog}(n)$$

then level-ℓ SoS algorithm strongly recovers the order-p discrete spiked tensor model.

when $\ell = n^{\delta}$ for $\delta \in (0, 1)$ the bound interpolates between λ_{SOS} and λ_{MLE} :

$$\lambda \gtrsim \begin{cases} 1 = \lambda_{MLE} & \delta \approx 1 \\ n^{\frac{p-2}{4}} = \lambda_{SOS} & \delta \approx 0 \end{cases}$$

Tensor unfolding

Algorithm [Montanari, Richard '14] :

1. reshape Y into matrix $Y \in \mathbb{R}^{n^q \times n^{p-q}}$
2. $Y \leftarrow$ leading left singular vector of Y; update p and q
3. iterate until $p = 1$

Theorem (Ben Arous, Huang, Huang '21)

If $\lambda \geq (1+\varepsilon) n^{\frac{p-2}{4}}$ where $\varepsilon \geq n^{-\frac{1}{3}-c} q$ then output of a step of tensor unfolding strongly correlates with the signal (leading to recovery). If $\varepsilon < n^{-\frac{1}{3}-c} q$ then output of a step of tensor unfolding is not correlated with the signal.

Kikuchi hierarchy

- For $E \subset [n] = \{1, \ldots, n\}$ s.t. $|E| = p$, the entry of the observed tensor Y_E

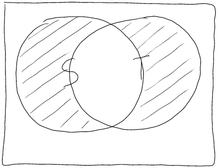
Kikuchi hierarchy

- For $E \subset [n] = \{1, \dots, n\}$ s.t. $|E| = p$, the entry of the observed tensor Y_E
- $\ell \in [\frac{p}{2}, n - \frac{p}{2}]$

Kikuchi hierarchy

- For $E \subset [n] = \{1, \ldots, n\}$ s.t. $|E| = p$, the entry of the observed tensor Y_E
- $\ell \in [\frac{p}{2}, n - \frac{p}{2}]$
- Symmetric $M \in \mathbb{R}^{\binom{n}{\ell} \times \binom{n}{\ell}}$ indexed by $S \subset [n]$ of $|S| = \ell$

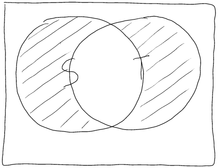
$$M_{S,T} = \begin{cases} Y_{S \triangle T} & \text{if } |S \triangle T| = p \\ 0 & \text{if } |S \triangle T| \neq p \end{cases}$$



Kikuchi hierarchy

- For $E \subset [n] = \{1, \ldots, n\}$ s.t. $|E| = p$, the entry of the observed tensor Y_E
- $\ell \in [\frac{p}{2}, n - \frac{p}{2}]$
- Symmetric $M \in \mathbb{R}^{\binom{n}{\ell} \times \binom{n}{\ell}}$ indexed by $S \subset [n]$ of $|S| = \ell$

$$M_{S,T} = \begin{cases} Y_{S \triangle T} & \text{if } |S \triangle T| = p \\ 0 & \text{if } |S \triangle T| \neq p \end{cases}$$



Algorithm [Wein, El Alaoui, Moore '19] :

- Compute leading eigenvector v of M
- Form symmetric $V \in \mathbb{R}^{n \times n}$ s.t.
  $$V_{ij} = \begin{cases} \frac{1}{2} \sum_{S \triangle T = \{i,j\}} v_S v_T & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$$
- Output leading eigenvector of V

Complexity result

Theorem (Wein, El Alaoui, Moore '19)

If the level of Kikuchi hierarchy $\ell = O(n)$ and

$$\lambda \gg \left(\frac{n}{\ell}\right)^{\frac{p-2}{4}} \sqrt{\log(n)}$$

then level-ℓ Kikuchi algorithm strongly recovers the order-p discrete spiked tensor model.

when $\ell = n^{\delta}$ for $\delta \in (0, 1)$ the bound interpolates between λ_{SOS} and λ_{MLE} again

Local search + Restarts

Theorem (Ben Arous, Gheissari, Jagannath '19)

If $\lambda \geq n^{\frac{p-2}{2}}$ then Gradient and Langevin dynamics recover signal in finite time with high probability. If $\lambda < n^{\frac{p-2}{2}}$ then the time is at least exponential.

Local search + Restarts

Theorem (Ben Arous, Gheissari, Jagannath '19)

If $\lambda \geq n^{\frac{p-2}{2}}$ then Gradient and Langevin dynamics recover signal in finite time with high probability. If $\lambda < n^{\frac{p-2}{2}}$ then the time is at least exponential.

Algorithm with restarts :

1. Run r independent copies of the dynamics
2. Output the best result

Local search + Restarts

Theorem (Ben Arous, Gheissari, Jagannath '19)

If $\lambda \geq n^{\frac{p-2}{2}}$ then Gradient and Langevin dynamics recover signal in finite time with high probability. If $\lambda < n^{\frac{p-2}{2}}$ then the time is at least exponential.

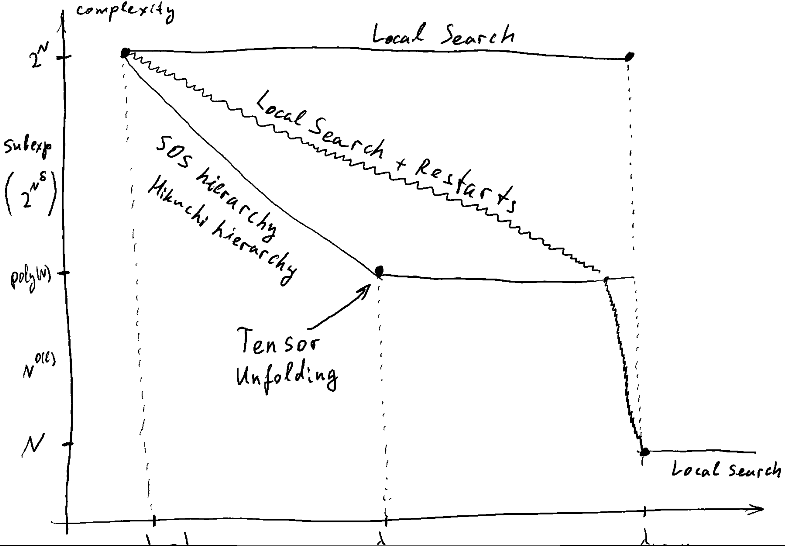
Algorithm with restarts :

1. Run r independent copies of the dynamics
2. Output the best result

Conjecture

- *If $\lambda \gg \frac{n^{\frac{k-2}{2}}}{(\beta \log n)^{k-2}}$ then $r = n^{\beta}$ is enough to strongly recover the order-p spiked tensor model*
- *If $\lambda \gg n^{(1-\beta)\frac{k-2}{2}}$ then $r = e^{n^{\beta}}$ is enough to strongly recover the order-p spiked tensor model (interpolates λ_{MLE} and λ_{Local})*

Performance of Restarts



Matrix sensing problem

- Order $p = 2$
- Low noise : $\lambda \approx \infty$

$$Y = \mathcal{A}\left(\sum_{i=1}^{r} u_i u_i^\top\right)$$

The problem is still NP-hard

Topological complexity

- $f : \mathcal{X} \to \mathbb{R}$
- sublevel (α) : $\{(x, y) | f(x) \leq y \leq \alpha\}$

Classical definition from convex optimization :

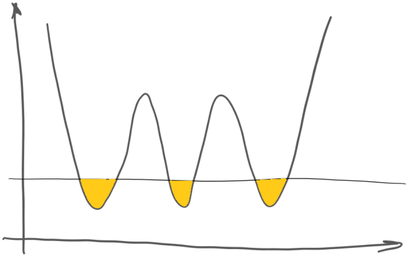
- Complexity is low if the number of connected components in sublevel (α) is 0 or 1.

Definition for non-convex optimization :

- Complexity is low if the number of connected components in sublevel (α) monotonically decreasing on $[\inf_{x \in \mathcal{X}} f(x), \infty]$

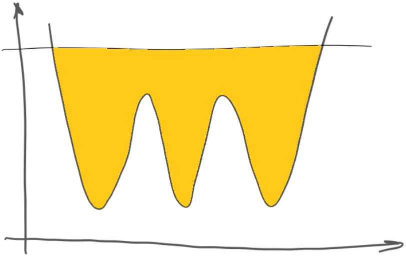
Topological complexity

- $f : \mathcal{X} \to \mathbb{R}$
- sublevel (α) : $\{(x, y) | f(x) \leq y \leq \alpha\}$
- Complexity is low if the number of connected components in sublevel (α) monotonically decreasing on $[\inf_{x \in \mathcal{X}} f(x), \infty]$



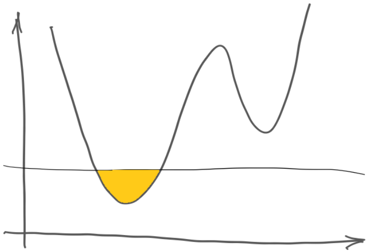
Topological complexity

- $f : \mathcal{X} \to \mathbb{R}$
- sublevel (α) : $\{(x, y) | f(x) \leq y \leq \alpha\}$
- Complexity is low if the number of connected components in sublevel (α) monotonically decreasing on $[\inf_{x \in \mathcal{X}} f(x), \infty]$



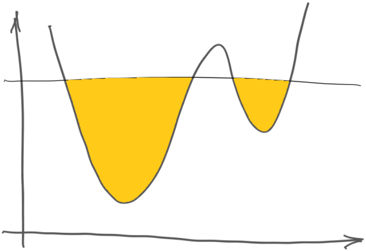
Topological complexity

- $f : \mathcal{X} \to \mathbb{R}$
- sublevel (α) : $\{(x, y) | f(x) \leq y \leq \alpha\}$
- Complexity is low if the number of connected components in sublevel (α) monotonically decreasing on $[\inf_{x\in\mathcal{X}} f(x), \infty]$



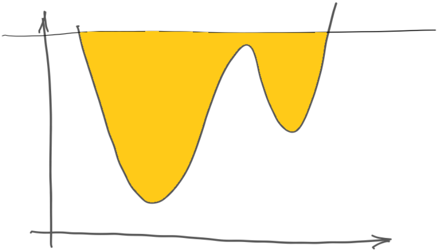
Topological complexity

- $f : \mathcal{X} \to \mathbb{R}$
- sublevel (α) : $\{(x, y) | f(x) \le y \le \alpha\}$
- Complexity is low if the number of connected components in sublevel (α) monotonically decreasing on $[\inf_{x \in \mathcal{X}} f(x), \infty]$

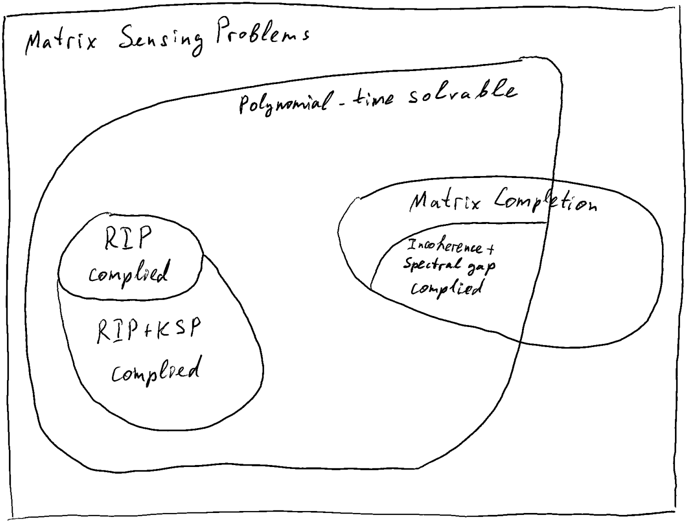


Topological complexity

- $f : \mathcal{X} \to \mathbb{R}$
- sublevel (α) : $\{(x, y) | f(x) \leq y \leq \alpha\}$
- Complexity is low if the number of connected components in sublevel (α) monotonically decreasing on $[\inf_{x \in \mathcal{X}} f(x), \infty]$



Structure + no noise = low complexity



Restricted Isometry Property

Definition

Lin map $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ satisfies δ_r-RIP for some $\gamma > 0$

$$(1 - \delta_r)\|X\|_F^2 \leq \gamma\|\mathcal{A}(X)\|_2^2 \leq (1 + \delta_r)\|X\|_F^2$$

holds for all X s.t. $\text{rank}(X) \leq r$.

Theorem (Candes, Plan '10)

If $\mathcal{A}$ is represented with a matrix of i.i.d $\mathcal{N}(0,1)$ random variables with $m = O(\frac{nr}{\delta_r^2})$, it satisfies δ_r-RIP w.h.p.

Recovery under RIP

Convexefication algorithm : $\min_{X \in \mathbb{R}^{n \times n}} \left\{ \|X\|_* \middle| \mathcal{A}(X) = Y \right\}$

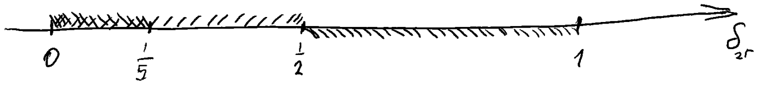
Theorem (Recht, Fazel, Parrilo '08)

Exact recovery under $\delta_r < 1/10$

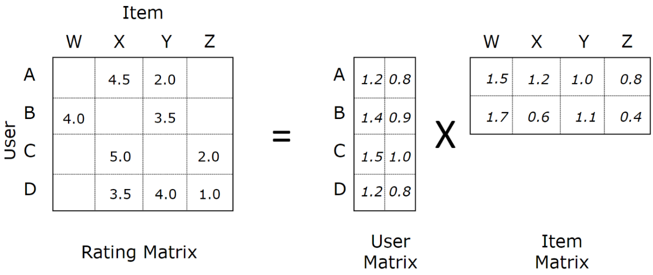
Local algorithm : $\min_{X \in \mathbb{R}^{n \times r}} \|\mathcal{A}(XX^T) - Y\|$

Theorem (Bhojanapalli, Neyshabur, Srebro '16, Zhang, Sojoudi, Lavaei '19)

- *If* $\delta_{2r} < 1/5$, *no second-order critical point with positive value (spurious).*
- *If* $r = 1$ *and* $\delta_2 < 1/2$, *no spurious second-order critical point.*



Matrix completion



Matrix completion problem

For some ordered $\Omega = \{i_k, j_k\}_{k=1}^{m} \subset [n] \times [n]$ define sensing operator of matrix completion $\mathcal{A}_\Omega$ s.t.

$$\mathcal{A}_\Omega(X)_k = X_{i_k j_k}$$

- Ω can be viewed as the adjacency matrix of a graph G
- RIP property is not satisfied for any δ and r

Solution of matrix completion

Convexefication algorithm : $\min_{X \in \mathbb{R}^{n \times n}} \left\{ \|X\|_* \middle| \mathcal{A}(X) = Y \right\}$
Assumptions (spectral gap + incoherence) :

- G is a d-regular graph with $\sigma_1(\Omega) = d$ and $\sigma_2(\Omega) \leq C\sqrt{d}$
- $\|u_i\|^2 \leq \frac{\mu_0 r}{n}$
- $\|\frac{n}{d} \sum_{i \in S} u_i u_i^\top - I\|^2 \leq \delta_d$ for all $S \subset [n]$ s.t. $|S| = d$

Solution of matrix completion

Convexefication algorithm : $\min_{X \in \mathbb{R}^{n \times n}} \left\{ \|X\|_* \middle| \mathcal{A}(X) = Y \right\}$
Assumptions (spectral gap + incoherence) :

- G is a d-regular graph with $\sigma_1(\Omega) = d$ and $\sigma_2(\Omega) \leq C\sqrt{d}$
- $\|u_i\|^2 \leq \frac{\mu_0 r}{n}$
- $\|\frac{n}{d} \sum_{i \in S} u_i u_i^\top - I\|^2 \leq \delta_d$ for all $S \subset [n]$ s.t. $|S| = d$

Proposition

$\sigma_2(\Omega) \leq C\sqrt{d}$ satisfied with high probability if G is uniformly sampled d-regular graph

Solution of matrix completion

Convexefication algorithm : $\min_{X\in\mathbb{R}^{n\times n}}\left\{\|X\|_*\middle|\mathcal{A}(X)=Y\right\}$
Assumptions (spectral gap + incoherence) :

- G is a d-regular graph with $\sigma_1(\Omega)=d$ and $\sigma_2(\Omega)\leq C\sqrt{d}$
- $\|u_i\|^2\leq\frac{\mu_0 r}{n}$
- $\|\frac{n}{d}\sum_{i\in S}u_iu_i^\top - I\|^2\leq\delta_d$ for all $S\subset[n]$ s.t. $|S|=d$

Proposition

$\sigma_2(\Omega)\leq C\sqrt{d}$ satisfied with high probability if G is uniformly sampled d-regular graph

Theorem (Bhojanapalli, Jain ’14, Ge, Lee, Ma ’16)

If $d\gtrsim\mu_0^2r^2$ and $\delta_d\leq\frac{1}{6}$ then nuclear norm minimization recovers U exactly.

Dense noise

$$Y = \mathcal{A}\left(\sum_{i=1}^{r} u_i u_i^\top\right) + w$$

Local search algorithm : $\min_{X \in \mathbb{R}^{n \times r}} \|\mathcal{A}(XX^T) - Y\|$

Theorem (Bhojanapalli, Neyshabur, Srebro '16)

w is i.i.d $\mathcal{N}(0, \sigma^2)$, $\mathcal{A}$ satisfies δ_{2r}-RIP with $\delta_2 r < 1/10$ then w.p. $1 - \frac{10}{n^2}$

$$X - \textit{local minimum} \Rightarrow \|XX^\top - UU^\top\|_F \leq 20\sigma\sqrt{\frac{\log n}{m}}$$

Sparse noise

Robust Quadratic Regression :

$$Y = \mathcal{A}(uu^\top) + w$$

- Rank $r = 1$
- $\|w\|_0 = \ell$

Theorem (M., Madani, Lavaei '20)

Penalised directed convex relaxation recovers the signal exactly w.h.p. under Gaussian data and $\ell \sim \sqrt{n}$

Conclusion

- Two important special instances of Tensor Sensing has been extensively studied : Tensor PCA and Matrix Sensing
- Both instances can be attacked with convexefication or local search techniques. The guarantees on their performance are similar
- Generalization to Tensor Sensing would be interesting for theoretical and practical reasons

Questions ?

Thank you for your attention !

Questions ?