Learning in graphs with semi-definite programming: statistical bounds based on fixed point analysis and excess risk curvature

Stephane Chretien (University of Lyon 2)

Many statistical learning problems on graphs have recently been shown to be amenable to Semi-Definite Programming (SDP), with community detection and clustering in Gaussian mixture models as the most striking instances Javanmard et al. (2016). Given the growing range of applications of SDP-based techniques to machine learning problems, and the rapid progress in the design of efficient algorithms for solving SDPs, an intriguing question is to understand how the recent advances from empirical process theory and Statistical Learning Theory can be leveraged for providing a precise statistical analysis of SDP estimators.

In this talk, we borrow cutting edge techniques and concepts from the Learning Theory literature, such as fixed point equations and excess risk curvature arguments, which yield general estimation and prediction results for a wide class of SDP estimators for estimation problems pertaining to gaphs. From this perspective, we revisit some classical results in community detection from Guédon and Vershynin (2016) and Fei and Chen (2019b), and we obtain statistical guarantees for SDP estimators used in signed clustering, angular group synchronization (for both multiplicative and additive models) and MAX-CUT.

Learning in graphs with semi-definite programming

statistical bounds based on fixed point analysis and excess risk curvature

Stéphane Chrétien
(joint with Mihai Cucuringu, Guillaume Lecué and Lucie Neirac)

Laboratoire ERIC,
University of Lyon 2

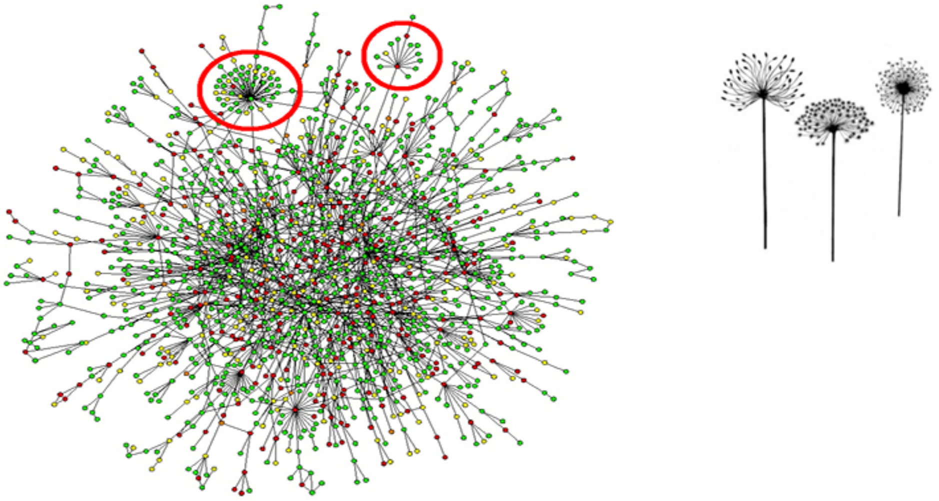
March 13, 2023

Part I

BACKGROUND

REAL VS RANDOM GRAPHS

EXAMPLES WITH LOCAL OR GLOBAL FEATURES



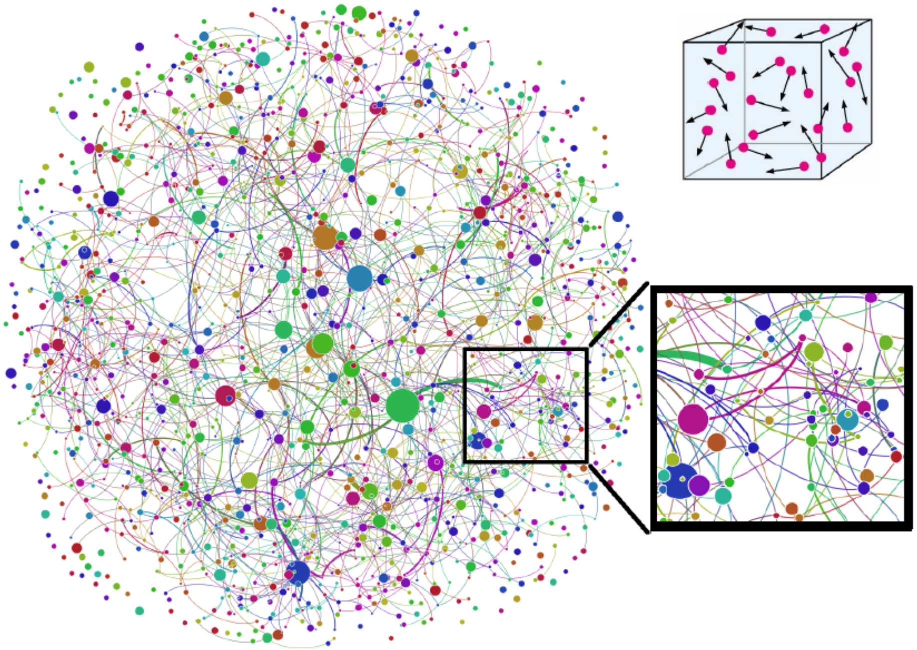
Protein interaction network
[A.-L. Barabási & Z. Oltvai, Nature Reviews Genetics 5, 101–113, Feb. 2004]

Figure. credit: Vershynin

Some structures are local.

Real vs random graphs

Examples with local or global features



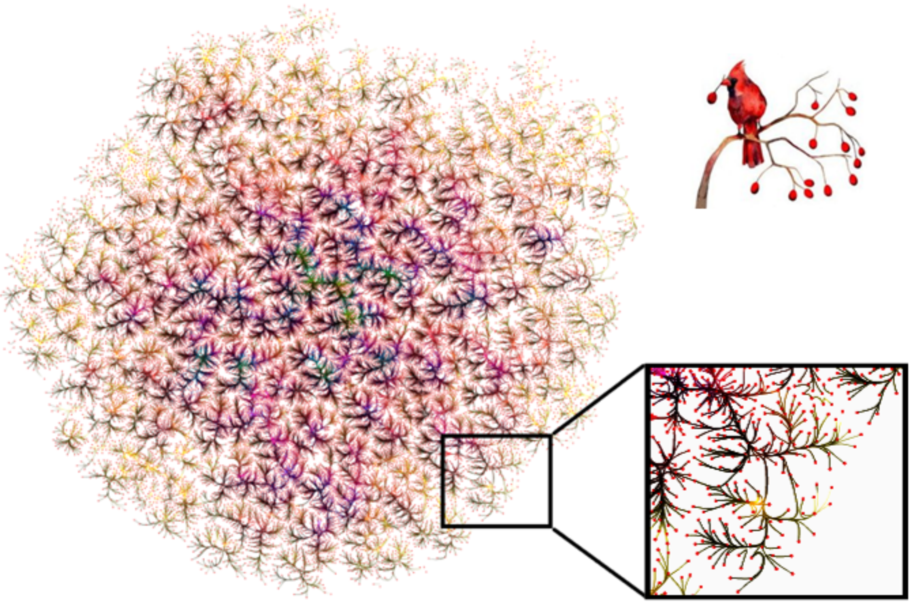
Collaboration network of economists

(AER, JPE, Econometrica, RES, QJE. www.cloudycnen.net)

Figure. credit: Vershynin

Real vs random graphs

Random graphs and their features



The Internet

(C. Hurter et al., Eurographics Conference on Visualization 2012)

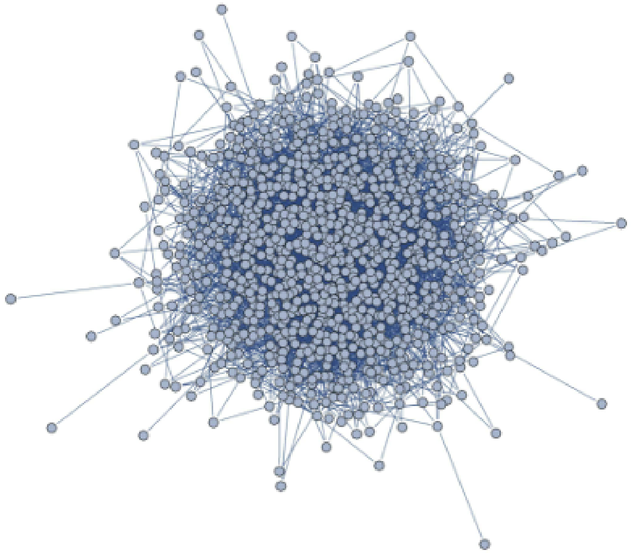
Figure. credit: Vershynin

Real vs random graphs

Random graphs and their features

Let us compare with stochastic models ! The first is the Erdoes-Renyi random graph.

$G(n, p)$ with $n = 1000$, $p = 0.00095$



(A. Novozhilov's course in Mathematics of Networks, NDSU)

Figure. credit: Vershynin

Real vs random graphs

Random graphs and their features

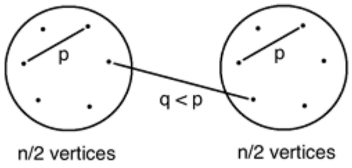
Let us compare with stochastic models ! A better model for clustered data is the Stochastic Block Model.

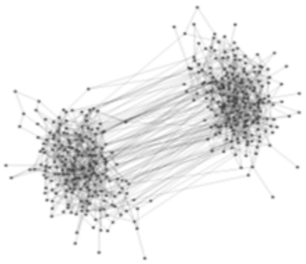
Edges are still independent, but can have different probabilities p_{ij}.

- Allows to model networks with structure = communities (clusters).

Example: Stochastic block model with two communities $G(n, p, q)$:

- Edges within each community: probability p;
- across communities: probability $q < p$.





Real vs random graphs

Random graphs and their features

Let us compare with stochastic models ! A better model for clustered data is the Stochastic Block Model.

Example: Stochastic block model with multiple communities $G(n, (p_{k,k'}))$:

- Edges within each community: probability p;
- across communities: probability $q < p$.

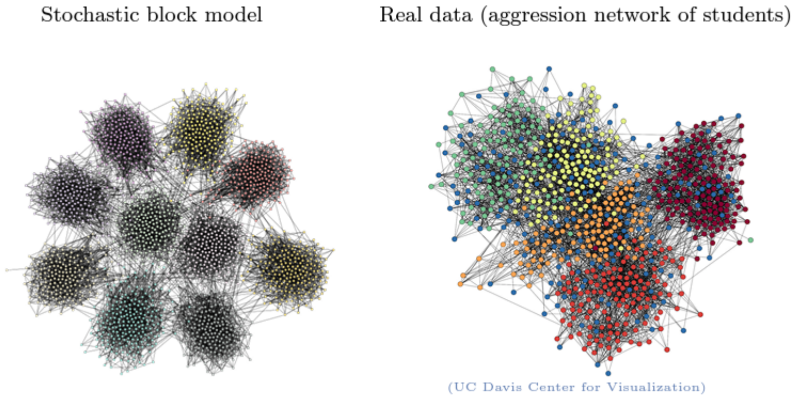
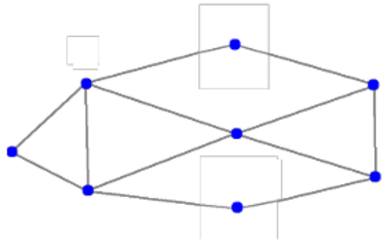

Stochastic block model
Real data (aggression network of students)
(UC Davis Center for Visualization)

Figure. credit: Vershynin

Concentration of Random Graphs

The adjacency matrix approach

Adjacency matrix A:



$$\longrightarrow \quad A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Figure. Recovering structures = communities

For inhomogeneous Erdös-Rényi model:

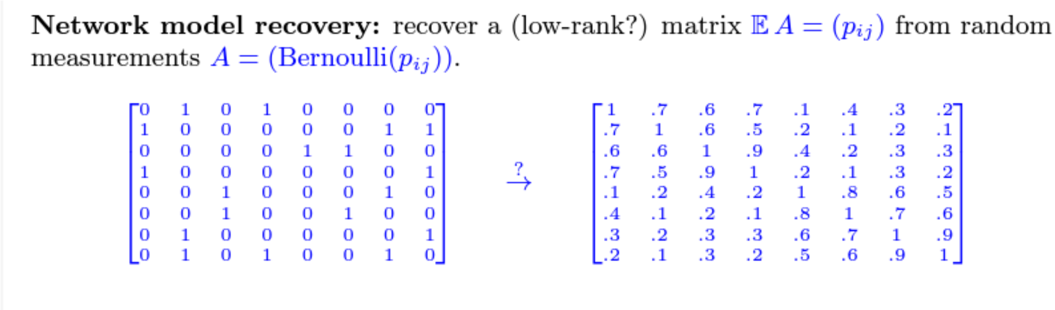
$$A = \left(\text{Bernoulli}\left(p_{ij}\right)\right), \quad \mathbb{E}A = \left(p_{ij}\right) \tag{1}$$

Concentration of Random Graphs

The adjacency matrix approach

Model Recovery Problem:

- Observe A
- recover $\mathbb{E}A$.



Network model recovery: recover a (low-rank?) matrix $\mathbb{E}\,A = (p_{ij})$ from random measurements $A = (\text{Bernoulli}(p_{ij}))$.

$$\begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} \overset{?}{\to} \begin{bmatrix} 1 & .7 & .6 & .7 & .1 & .4 & .3 & .2 \\ .7 & 1 & .6 & .5 & .2 & .1 & .2 & .1 \\ .6 & .6 & 1 & .9 & .4 & .2 & .3 & .3 \\ .7 & .5 & .9 & 1 & .2 & .1 & .3 & .2 \\ .1 & .2 & .4 & .2 & 1 & .8 & .6 & .5 \\ .4 & .1 & .2 & .1 & .8 & 1 & .7 & .6 \\ .3 & .2 & .3 & .3 & .6 & .7 & 1 & .9 \\ .2 & .1 & .3 & .2 & .5 & .6 & .9 & 1 \end{bmatrix}$$

Figure. Model recovery (from a talk by Vershynin)

Concentration of Random Graphs

The adjacency matrix approach

Let us consider the simpler problem of identifying the structure of a random graph from a unique observation of the graph.

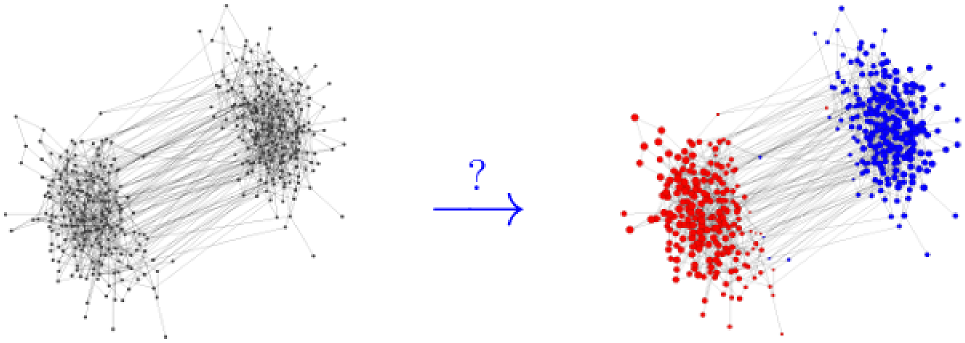


Figure. Recovering structures = communities

Concentration of Random Graphs

The adjacency matrix approach

Eigenvectors reveal the latent structure !

- If concentration (possibly after regularization) ⇒

$$A \approx \mathbb{E}A$$

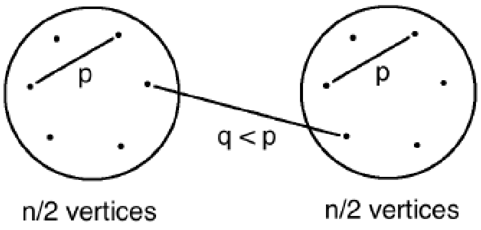
- The Davis-Kahan theorem ⇒ eigenvectors satisfy

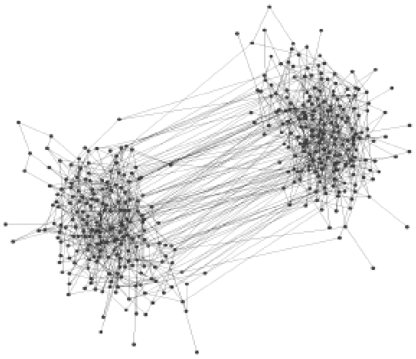
$$v_i(A) \approx v_i(\mathbb{E}A)$$

AND: Eigenvectors $v_i(\mathbb{E}A)$ carry information about network structure.

Concentration of Random Graphs

The adjacency matrix approach





$$\mathbb{E}A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right] \text{ has rank 2;} \quad v_1(\mathbb{E}A) = \left[\begin{array}{c} 1 \\ 1 \\ \hline 1 \\ 1 \end{array}\right], \quad v_2(\mathbb{E}A) = \left[\begin{array}{r} 1 \\ 1 \\ \hline -1 \\ -1 \end{array}\right]$$

$v_2(\mathbb{E}A)$ encodes community structure $\Rightarrow v_2(A)$ encodes the structure, too.

Concentration of Random Graphs

The adjacency matrix approach

How will the Davis Kahan theorem help ?

$$A \approx \mathbb{E}A \text{ in a certain way} \tag{2}$$

implies

$$\text{eigenstructure}(A) \approx \text{eigenstructure}(\mathbb{E}A) \tag{3}$$

Theorem 1 (Davis-Kahan $\sin(\Theta)$ theorem)

Let $A = E_0 A_0 {E_0}^* + E_1 A_1 {E_1}^*$ and $A + H = F_0 \Lambda_0 {F_0}^* + F_1 \Lambda_1 {F_1}^*$ be symmetric matrices with $[E_0, E_1]$ and $[F_0, F_1]$ orthogonal.
If the eigenvalues of A_0 are contained in an interval (a, b), and the eigenvalues of Λ_1 are excluded from the interval $(a - \delta, b + \delta)$ for some $\delta > 0$, then

$$\|{F_1}^* E_0\| \leq \frac{\|{F_1}^* H E_0\|}{\delta} \tag{4}$$

for any unitarily invariant norm $\| \cdot \|$.

Concentration of Random Graphs

The adjacency matrix approach

Question: Do random graphs concentrate near their "expected" graphs?

Concentration of Random Graphs

The adjacency matrix approach

Theorem 2

An inhomogeneous Erdös-Rényi random graph $G\left(n, \left(p_{ij}\right)\right)$ with expected degrees $np_{ij} \sim d$ with $d \gtrsim \log n$ concentrates:

$$\|A - \mathbb{E}A\| \lesssim \sqrt{d} \quad \textit{w.h.p. while } \|\mathbb{E}A\| \sim d.$$

Proof. Simple concentration of

$$x^\top (A - \mathbb{E}A) y$$

for fixed x and y. Then, complicated union bound over x, y. □.

Weaker earlier results by Furedi and Komlos (80's) with $d \gtrsim \log^4 n$. Oliveira also obtained a result in this spirit in 2010, with $\|A - \mathbb{E}A\| \lesssim \sqrt{d \log n}$, using the matrix Bernstein inequality (Tropp).

Concentration of Random Graphs

The adjacency matrix approach

Observation: A random graph $G(n, p)$ with expected degrees $d = np \ll \log n$ does not concentrate !

$$\|A - \mathbb{E}A\| \gg \|\mathbb{E}A\|$$

What is wrong with sparse graphs?

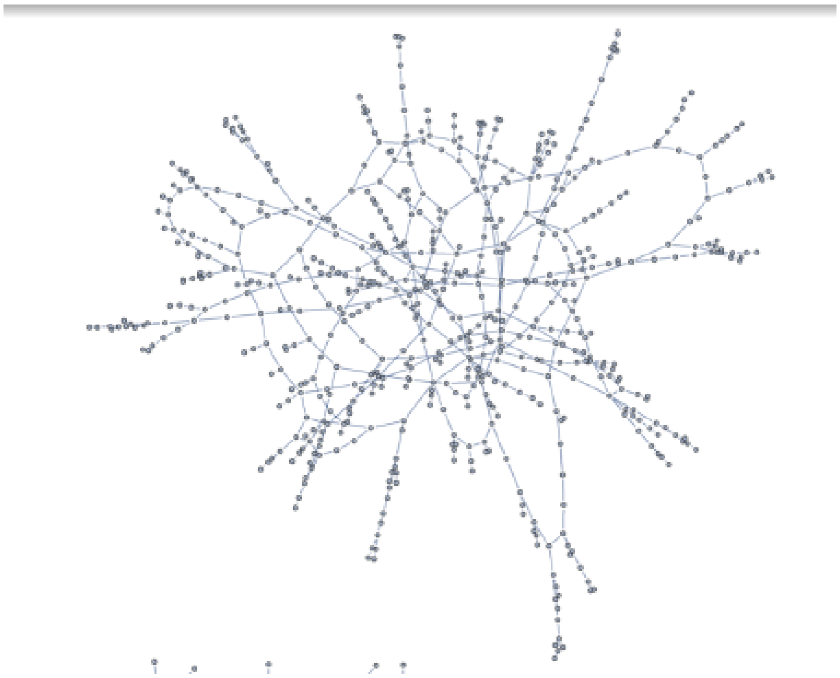
The degrees are wild, do not concentrate near d anymore. High-degree vertices blow up $\|A\|$: some columns of A are too large.

Concentration of Random Graphs

The adjacency matrix approach

Observation: A random graph $G(n, p)$ with expected degrees $d = np \ll \log n$ does not concentrate !

$$\|A - \mathbb{E}A\| \gg \|\mathbb{E}A\|$$



Concentration of Random Graphs

The adjacency matrix approach

Regularization and concentration:

Inhomogeneous E-R random graph with $d = \max np_{ij}$.

- Regularize vertices with degrees $> 2d$:
  - make all degrees $\leq 2d$ by reducing the weights of edges arbitrarily.

Theorem 3 (Le-Levina-Vershynin (2015))

The adjacency matrix A' of the regularized graph concentrates:

$$\|A' - \mathbb{E}A\| \lesssim \sqrt{d} \quad \textit{w.h.p.}$$

Structure estimation

The spectral approach

Recall that : Eigenvectors reveal the latent structure !

Concentration (possibly after regularization) ⇒

$$A \approx \mathbb{E}A$$

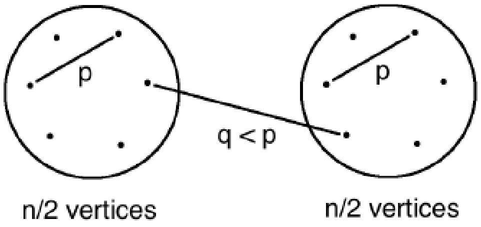
The Davis-Kahan theorem ⇒ eigenvectors satisfy

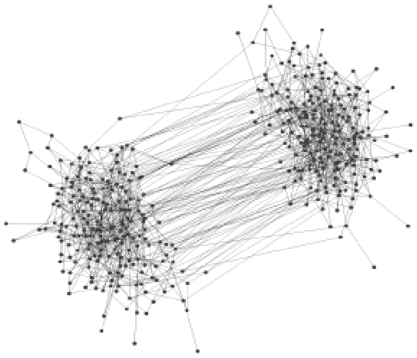
$$v_i(A) \approx v_i(\mathbb{E}A)$$

Moreover : Eigenvectors $v_i(\mathbb{E}A)$ carry information about network structure.

Structure estimation

The spectral approach





$$\mathbb{E}A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right] \text{ has rank 2;} \quad v_1(\mathbb{E}A) = \left[\begin{array}{c} 1 \\ 1 \\ \hline 1 \\ 1 \end{array}\right], \quad v_2(\mathbb{E}A) = \left[\begin{array}{r} 1 \\ 1 \\ \hline -1 \\ -1 \end{array}\right]$$

$v_2(\mathbb{E}A)$ encodes community structure $\Rightarrow$ $v_2(A)$ encodes the structure, too.

Structure estimation

The spectral approach

Spectral Clustering Algorithm:

- given a graph with adjacency matrix A,
  - Compute the second leading eigenvector of A;
  - Recover communities based on the signs of its coefficients.

$$\mathbb{E}A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right] \text{ has rank 2;} \quad v_1(\mathbb{E}A) = \left[\begin{array}{c} 1 \\ 1 \\ \hline 1 \\ 1 \end{array}\right], \quad v_2(\mathbb{E}A) = \left[\begin{array}{r} 1 \\ 1 \\ \hline -1 \\ -1 \end{array}\right]$$

STRUCTURE ESTIMATION

THE SPECTRAL APPROACH

Corollary 1 (Community Detection)

Consider the stochastic block model $G(n, p, q)$ with $p = a/n$ and $q = b/n$. Suppose

$$(a - b)^2 \geq C_\varepsilon (a + b)$$

Then the regularized spectral clustering algorithm recovers communities up to εn misclassified vertices, and with high probability.

Proof: This is a consequence of the concentration result of [Le-Levina-Vershynin (2015)], combined with the Davis Kahan Theorem. □

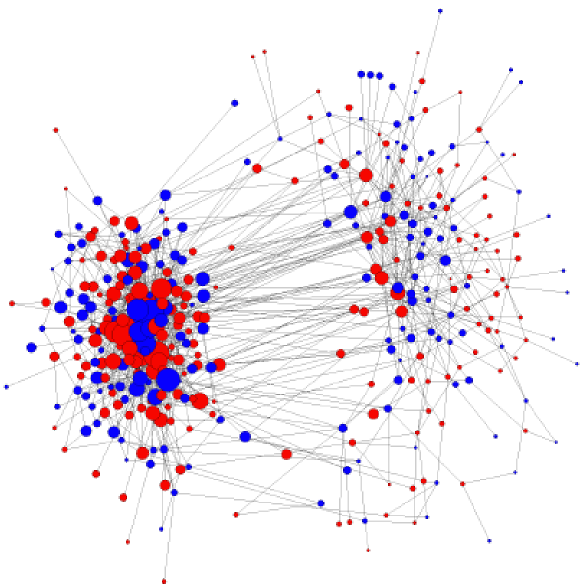
Detection threshold

- The condition on $(a-b)^2 \geq C_\varepsilon(a+b)$ is optimal up to C_ε,
  - $C_\epsilon \to \infty$ as $\epsilon \to 0$.
- No algorithm can succeed if
$$(a-b)^2 \leq 2(a+b).$$
- There are algorithms that do better than random guess if
$$(a-b)^2 > 2(a+b).$$

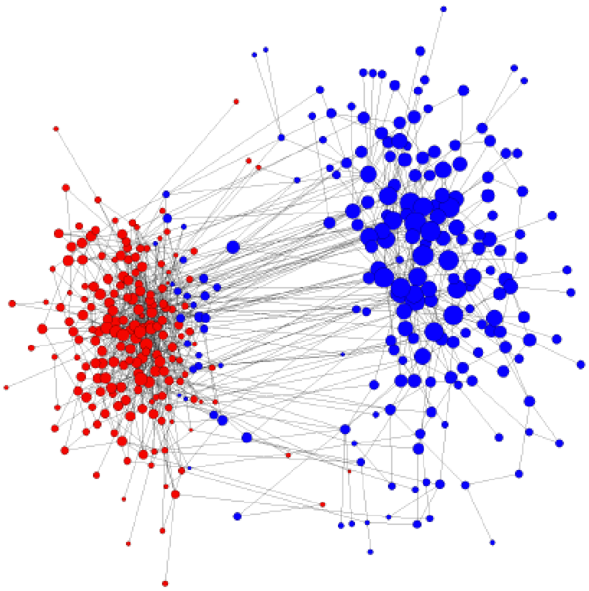
This is work by Mossel-Neeman-Sly (2013-14) and Massoulié (2013).

Structure estimation

The spectral approach



Without regularization



With regularization

Structure estimation

The spectral approach

On a graph, the discrete Laplacian is the $n \times n$ matrix

$$\Delta := I - D^{-1/2} A D^{-1/2}$$

where D is the diagonal matrix with the degrees on the diagonal.

Do Laplacians concentrate as Adjacency matrices ?

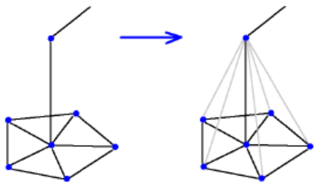
- For dense graphs (expected degrees $d \gtrsim \log n$) , Laplacian concentrates.
- For sparse graphs $d \ll \log n$, fails to concentrate.
    - What's wrong?
        - Low-degree vertices: isolated vertices, trees. (They get overheated.)

Structure estimation

The spectral approach

How to regularise for helping Laplacians to concentrate ?

- Connect low-degree vertices to the rest of the graph by light weighted edges; bring up all degrees to $\sim d$.



$$\Longleftrightarrow \qquad A_{ij} \mapsto A_{ij} + d/n$$

Structure estimation

The spectral approach

Theorem 4 (Concentration of Laplacians)

The Laplacian Δ' of the regularized graph concentrates !

$$\|\Delta' - \mathbb{E}\Delta'\| \lesssim \frac{1}{\sqrt{d}} \quad \textit{while} \quad \|\Delta'\| \sim 1$$

Proof: Deduced from concentration of regularized adjacency matrices. *Box*

Application to community detection: use the 2nd eigenvector of the Laplacian.

- Theoretical performance:
  - same as for adjacency
- empirically even better.

Part II

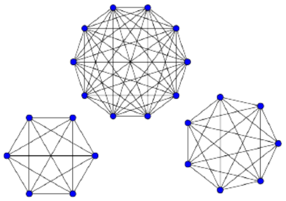
The Semi-Definite Programming Approach

Introduction

- Strongest community structure: union of cliques.
- How to fit?
  - Maximize correlation between the network and a union of cliques.
- Optimization:

$$\max_{Z \in \left\{\begin{array}{c}\text{adjacency matrices of a union of}\\ \text{cliques with } k \text{ edges}\end{array}\right\}} \langle A, Z \rangle$$

  - where A = adjacency matrix of the network,



$$Z = \begin{bmatrix} 1 & 1 & 1 & 1 & & & & & \\ 1 & 1 & 1 & 1 & & & & & \\ 1 & 1 & 1 & 1 & & & & & \\ 1 & 1 & 1 & 1 & & & & & \\ & & & & 1 & 1 & 1 & & \\ & & & & 1 & 1 & 1 & & \\ & & & & 1 & 1 & 1 & & \\ & & & & & & & 1 & 1 \\ & & & & & & & 1 & 1 \end{bmatrix}$$

Introduction

- This is equivalent to

$$\max_Z \langle A, Z\rangle : \quad Z \in \{0,1\}^{n\times n} \quad \text{is block-diagonal}, \quad \sum_{ij} Z_{ij} = k. \tag{5}$$

Lemma 1

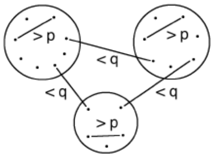
A matrix $Z \in \{0,1\}^{n\times n}$ is block diagonal $\Leftrightarrow$ Z is positive semidefinite.

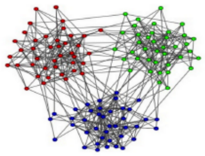
- A semidefinite (SDP) relaxation:

$$\max_Z \langle A, Z\rangle : \quad Z \in [0,1]^{n\times n} \text{ is positive semidefinite}, \ \sum_{ij} Z_{ij} = k.$$

Introduction

- General stochastic block model: $\forall$ many communities, $\forall$ connection probabilities p_{ij}, within communities $> p$; across communities $< q$. (Not necessarily low rank!)





Theorem 5 (Guedon and Vershynin)

Consider a general stochastic block model with $p = a/n$ and $q = b/n$. Suppose

$$(a-b)^2 \geq C_\varepsilon (a+b)$$

Then the SDP (with $k =$ number of edges) recovers communities up to εn misclassified vertices, and with high probability.

Introduction

- SemiDefinite Programming is a class of optimization problems which includes linear programming as a particular case and can be written as
  - the set of problems over symmetric (resp. Hermitian) positive semi-definite matrix variables,
  - with
    - linear cost function and
    - affine constraints,

  i.e. optimization problems of the form

$$\max_{Z \succeq 0} \left(\langle A, Z \rangle : \langle B_j, Z \rangle = b_j \text{ for } j = 1, \ldots, m \right), \tag{6}$$

where A, B_1, ..., B_m are given matrices.

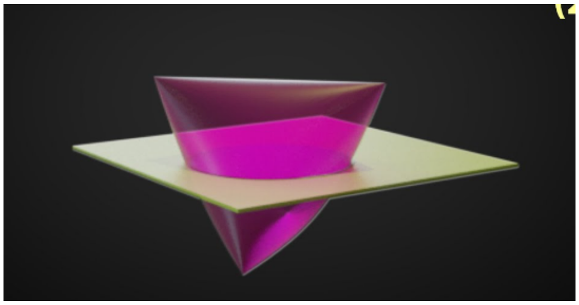
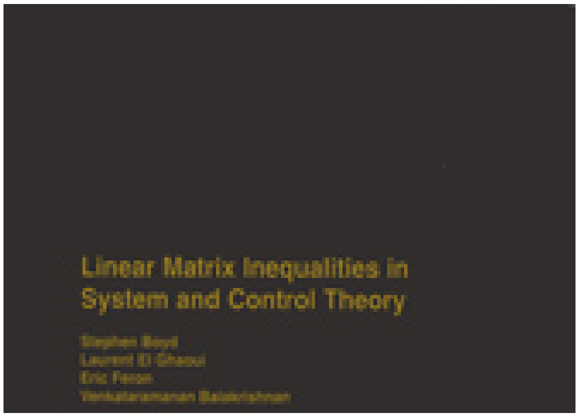


Figure. SemiDefinite Programming

Introduction

Early history

- Early use of Semi-Definite programming to statistics can be traced back to Scobey and Kabe 1978 and Fletcher 1981.
- In the same year, Shapiro used SDP in factor analysis Shapiro 1982.
- The study of the mathematical properties of SDP then gained momentum with the introduction of Linear Matrix Inequalities (LMI) and their numerous applications in control theory, system identification and signal processing.
- The book Boyd, El Ghaoui, et al. 1994 is the standard reference of these type of results, mostly obtained in the 90's.



Linear Matrix Inequalities in
System and Control Theory

Stephen Boyd
Laurent El Ghaoui
Eric Feron
Venkataramanan Balakrishnan

Introduction

The Goemans-Williamson SDP relaxation of Max-Cut and its legacy

- A notable turning point is the publication of Goemans and Williamson 1995 where SDP was shown to provide a 0.87 approximation to the NP-Hard problem known as Max-Cut.
  - The Max-Cut problem is a clustering problem on graphs which consists in finding two complementary subsets S and S^c of nodes such that the sum of the weights of the edges between S and S^c is maximal.

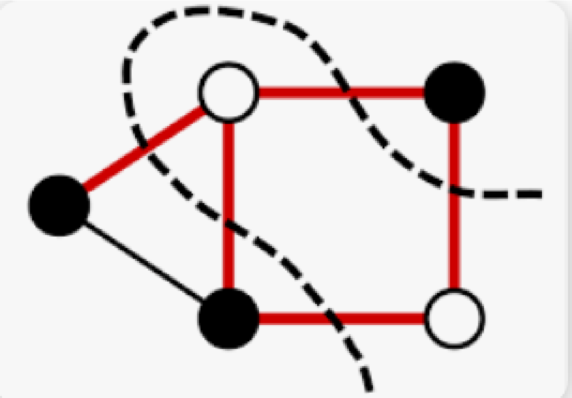


Figure. The Max-Cut problem

Introduction

The Goemans-Williamson SDP relaxation of Max-Cut and its legacy

- In Goemans and Williamson 1995, the authors approach this difficult combinatorial problem by using what is now known as the Goemans-Williamson *SDP relaxation* and use the Choleski factorization of the optimal solution to this SDP in order to produce a randomized scheme achieving the .87 bound in expectation !
- Moreover, this problem can be seen as one of the first prominent instances where the Laplacian of a graph is employed in order to provide an optimal bi-clustering in a graph and
  - certainly represents for a lot of people the first chapter of a long and fruitful relationship between clustering, embedding and Laplacians (but ... remember Delorme and Poljak !) .
- Other SDP schemes for approximating hard combinatorial problems are, to name a few, for the graph coloring problem Karger, Motwani, and Sudan 1998, for satisfiability problem Goemans and Williamson 1995; Goemans and Williamson 1994. These results were later surveyed in Lemaréchal, Nemirovskii, and Yurii Nesterov 1995; Goemans 1997 and Wolkowicz 1999.
- The randomized scheme introduced by Goemans and Williamson was then further improved in order to study more general Quadratically Constrained Quadratic Programmes (QCQP) in various references, most notably Nesterov 1997; Zhang 2000 and further extended in He et al. 2008.
- Many applications to signal processing are discussed in Olsson, Eriksson, and Kahl 2007, Ma 2010; one specific reduced complexity implementation in the form of an eigenvalue minimization problem and its application to binary least-squares recovery and denoising is presented in Chrétien and Corset 2009.

Introduction

The Goemans-Williamson SDP relaxation of Max-Cut and its legacy

- Applications of SDP to problems related with machine learning is more recent and probably started with the SDP relaxation of K-means in Peng and Xia 2005; Peng and Wei 2007 and later in Ames 2014.
- This approach was then further improved using a refined statistical analysis by Royer 2017 and Giraud and Verzelen 2018.
  - Similar methods have also been applied to community detection Hajek, Wu, and J. Xu 2016; Abbe, Bandeira, and Hall 2015 and for the weak recovery viewpoint, Guédon and Vershynin 2016.
- This last approach was also re-used via the kernel trick for the point cloud clustering Chrétien, Dombry, and Faivre to appear.
- Another incarnation of SDP in machine learning is the extensive use of nuclear norm-penalized least-squares costs as a surrogate for rank-penalization in low-rank recovery problems such as matrix completion in recommender systems, matrix compressed sensing, natural language processing and quantum state tomography; these topics are surveyed in Davenport and Romberg 2016.
- The problem of manifold learning was also addressed using SDP and is often mentioned as one of the most accurate approaches to the problem, let aside its computational complexity; see Weinberger, Packer, and Saul 2005; Weinberger and Saul 2006b; Weinberger and Saul 2006a; Hegde, Sankaranarayanan, and Baraniuk 2012. Connections with the design of fast converging Markov-Chains were also exhibited in Sun et al. 2006.

Part III

Analysis of SDP estimators using pixed point and curvature

Mathematical formulation of the problem

- The general problem we study can be stated as follows. Let A be a random matrix in $\mathbb{R}^{n\times n}$ and $\mathcal{C} \subset \mathbb{R}^{n\times n}$ be a constraint.

- The object that we want to recover, for instance, the community membership vector in community detection, is related to an *oracle* defined as

$$Z^* \in \operatorname*{arg\,max}_{Z\in\mathcal{C}} \langle \mathbb{E}A, Z\rangle, \tag{7}$$

  where

$$\langle A, B\rangle = \mathrm{Tr}(A\bar{B}^\top) = \sum A_{ij}\bar{B}_{ij} \text{ when } A, B \in \mathbb{C}^{n\times n}$$

  where $\bar{z}$ is the conjugate of $z \in \mathbb{C}$.

- We would like to estimate Z^*,
  - from which we can ultimately retrieve the object that really matters to us
    - (for instance, by considering a singular vector associated to the largest singular value of Z^*).

Mathematical formulation of the problem

- One way to go is to consider the following natural estimator of Z^* given by:

$$\hat{Z} \in \arg\max_{Z \in \mathcal{C}} \langle A, Z \rangle, \tag{8}$$

which is simply obtained by replacing the unobserved quantity $\mathbb{E}A$ in

$$Z^* \in \arg\max_{Z \in \mathcal{C}} \langle \mathbb{E}A, Z \rangle, \tag{9}$$

by the observed matrix A.

Mathematical formulation of the problem

- As pointed out above, in many situations, Z^* is not the object we want to estimate, but
  - there is a straightforward relation between Z^* and the object we seek to recover.
  - For instance, consider the community detection problem, where the goal is to recover the class community vector $x^* \in \{-1,1\}^n$ of n nodes.
    - Here, when $\mathcal{C}$ is well chosen, there is a close relation between Z^* and x^*, given by $Z^* = x^*(x^*)^\top$, e.g.
      $$x_i^{*^2} = 1 \Leftrightarrow x_i^* \in \{-1,1\}$$
      .
    - We therefore need a final step to estimate x^* using $\hat{Z}$, for instance, by letting $\hat{x}$ denote a top eigenvector of $\hat{Z}$, and then using the Davis-Kahan "sin-Theta" Theorem Davis and Kahan 1970; Yu, Wang, and Samworth 2015 to control the estimation of x^* by $\hat{x}$ from the one of Z^* by $\hat{Z}$.
- When the constraint $\mathcal{C}$ is of the form
  $$\mathcal{C} = \{Z \in \mathbb{R}^{n\times n} : Z \succeq 0, \langle Z, B_j\rangle = b_j, j = 1, \ldots, m\}$$
  where $B_1, \ldots, B_m \in \mathbb{R}^{n\times n}$ and $Z \succeq 0$ is notation for " Z is positive semidefinite", then
  $$\hat{Z} \in \underset{Z\in\mathcal{C}}{\arg\max}\ \langle A, Z\rangle,$$
  is a semidefinite programming (SDP) problem Boyd and Vandenberghe 2004

Goal of the paper

- The aim of the present work is to present a general approach to the study of the statistical properties of SDP-based estimators defined by

$$\hat{Z} \in \arg\max_{Z \in \mathcal{C}} \left\langle A, Z \right\rangle,$$

- In particular, using our framework, one is able to obtain new (non-asymptotic) rates of convergence or exact reconstruction properties for a wide class of estimators obtained as a solution of a semidefinite program like these.
- Specifically, our goal is to show that the solution to this problem can be analyzed in a statistical way when $\mathbb{E}A$ is only partially and noisily observed through A.
- Our analysis extends in a straightforward way to more complex sets $\mathcal{C}$ than SDP constraints !

MAIN GENERAL RESULTS FOR THE STATISTICAL ANALYSIS OF SDP ESTIMATORS

- From a statistical point of view, the task remains to estimate in the most efficient way the oracle Z^*.

- The point of view we will use to evaluate how far $\hat{Z}$ is from Z^* is coming from the Learning Theory literature.

  - We therefore see $\hat{Z}$ as an empirical risk minimizer (ERM) built on a single observation A, where the loss function is the linear one $Z \in \mathcal{C} \to \ell_Z(A) = -\langle A, Z \rangle$,

  - The most important requirement will be that the oracle Z^* is indeed the one minimizing the risk function $Z \in \mathcal{C} \to \mathbb{E}\ell_Z(A)$ over $\mathcal{C}$.

- Having this setup in mind, we can use all the machinery developed in Learning Theory to obtain rates of convergence for the ERM (here $\hat{Z}$) toward the oracle (here Z^*).

Main general results for the statistical analysis of SDP estimators

- We will introduce one key quantity that will be shown to control the rate of convergence of the ERM: a fixed point complexity parameter !
- This type of parameter carries all the statistical complexity of the problem, and even though it is usually easy to set up, its computation can be tedious since it requires to control, with large probability, the supremum of empirical processes indexed by "localized classes".
- Define this complexity fixed point related to the problem as

Definition 1

Let $0 < \Delta < 1$. *The fixed point complexity parameter at deviation* $1 - \Delta$ *is*

$$r^*(\Delta) = \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z \in \mathcal{C}: \left\langle \mathbb{E}A, Z^* - Z \right\rangle \leq r} \left\langle A - \mathbb{E}A, Z - Z^* \right\rangle \leq (1/2) r \right] \geq 1 - \Delta \right). \tag{10}$$

Main general results for the statistical analysis of SDP estimators

- Fixed point complexity parameters have been extensively used in Learning Theory since the introduction of the localization argument
  - When they can be computed, they are preferred to the (global) analysis developed by Chervonenkis and Vapnik (VC complexity).
  - VC analysis always yields slower rates since
    $$\sup_{Z\in\mathcal{C}} \left\langle A-\mathbb{E}A, Z-Z^*\right\rangle$$
    is an upper bound for $r^*(\Delta)$
    $$r^*(\Delta)=\inf\left(r>0:\mathbb{P}\left[\sup_{Z\in\mathcal{C}:\left\langle \mathbb{E}A,Z^*-Z\right\rangle\leq r}\left\langle A-\mathbb{E}A, Z-Z^*\right\rangle\leq(1/2)r\right]\geq 1-\Delta\right). \qquad (11)$$
    since $\{Z\in\mathcal{C}:\left\langle \mathbb{E}A,Z^*-Z\right\rangle\leq r\}\subset\mathcal{C}$.
- The gap between the two global and local analysis can be important since fast rates cannot be obtained using the VC approach, whereas the localization argument resulting in fixed points may yield fast rates of convergence or even exact recovery results.

Main general results for the statistical analysis of SDP estimators

- An example of a Vapnik-Chervonenkis's type of analysis of SDP estimators can be found in Guédon and Vershynin 2016 for the community detection problem.

- An improvement of the latter approach has been obtained in Fei and Chen 2019 thanks to a localization argument – even though it is not stated in these words

- On the other hand, Fixed point based analysis were proved to be optimal (in a minimax sense) when the noise $A - \mathbb{E}A$ is Gaussian Lecué and Mendelson 2013 and under mild conditions on the complexity of $\mathcal{C}$.

Main general results for the statistical analysis of SDP estimators

- Our main general statistical bound on SDP estimators is as follows.

Theorem 6

We assume that the constraint $\mathcal{C}$ is star-shaped in Z^*. Then, for all $0 < \Delta < 1$, with probability at least $1 - \Delta$, it holds true that $\langle \mathbb{E}A, Z^* - \hat{Z} \rangle \leq r^*(\Delta)$.

- Theorem 6 applies to any type of setup where an oracle Z^* is estimated by an estimator $\hat{Z}$ such as defined above.
- Its result shows that $\hat{Z}$ is almost a maximizer of the true objective function $Z \to \langle \mathbb{E}A, Z \rangle$ over $\mathcal{C}$ up to $r^*(\Delta)$.
  - In particular, when $r^*(\Delta) = 0$, $\hat{Z}$ is exactly a maximizer such as Z^* and, in that case, we can work with $\hat{Z}$ as if we were working with Z^* without any loss.
  - In this "exact reconstruction case", the information contained about A on $\mathbb{E}[A]$ is enough for knowing Z^* exactly.

Main general results for the statistical analysis of SDP estimators

- Let Ω^* be the event onto which for all $Z \in \mathcal{C}$ if $\langle \mathbb{E}A, Z^* - Z \rangle \leq r^*(\Delta)$ then
$$\langle A - \mathbb{E}A, Z - Z^* \rangle \leq (1/2) r^*(\Delta).$$
- By Definition of $r^*(\Delta)$, we have $\mathbb{P}[\Omega^*] \geq 1 - \Delta$.
- Let $Z \in \mathcal{C}$ be such that $\langle \mathbb{E}A, Z^* - Z \rangle > r^*(\Delta)$ and define Z' such that
$$Z' - Z^* = \big(r^*(\Delta)/\langle \mathbb{E}A, Z^* - Z \rangle\big)(Z - Z^*).$$
We have $\langle \mathbb{E}A, Z^* - Z' \rangle = r^*(\Delta)$ and $Z' \in \mathcal{C}$ because $\mathcal{C}$ is convex.
- Therefore, on the event Ω^*, $\langle A - \mathbb{E}A, Z' - Z^* \rangle \leq (1/2) r^*(\Delta)$ and therefore
$$\langle A - \mathbb{E}A, Z - Z^* \rangle \leq (1/2) \langle \mathbb{E}A, Z^* - Z \rangle.$$
- It therefore follows that on the event Ω^*, if $Z \in \mathcal{C}$ is such that $\langle \mathbb{E}A, Z^* - Z \rangle > r^*(\Delta)$ then
$$\langle A, Z - Z^* \rangle \leq (-1/2) \langle \mathbb{E}A, Z^* - Z \rangle < -r^*(\Delta)/2$$
which implies that $\langle A, Z - Z^* \rangle < 0$ and therefore Z does not maximize $Z \to \langle A, Z \rangle$ over $\mathcal{C}$.
- As a consequence, we necessarily have $\langle \mathbb{E}A, Z^* - \hat{Z} \rangle \leq r^*(\Delta)$ on the event Ω^* (which holds with probability at least $1 - \Delta$).

Main general results for the statistical analysis of SDP estimators

- The proof makes strong use of two important concepts originally introduced in Learning Theory, namely
  - the complexity of the problem comes from the one of the local subset

    $$\mathcal{C} \cap \{Z : \langle \mathbb{E}A, Z^* - Z \rangle \leq r^*(\Delta)\}$$

  - the "radius" $r^*(\Delta)$ of the localization is solution of a fixed point equation.

- In order to put this approach to work on concrete problems, we need to understand
  - the shape of the local subsets $\mathcal{C} \cap \{Z : \langle \mathbb{E}A, Z^* - Z \rangle \leq r\}, r > 0$ and
  - the maximal oscillations of the empirical process $Z \to \langle A - \mathbb{E}A, Z - Z^* \rangle$ indexed by these local subsets.

Main general results for the statistical analysis of SDP estimators

- Let us introduce an easier to compute proxy for $r^*(\Delta)$, defined as

  $$r^*_G(\Delta) = \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z \in \mathcal{C}: G(Z^* - Z) \leq r} \left\langle A - \mathbb{E}A, Z - Z^* \right\rangle \leq (1/2)r\right] \geq 1 - \Delta\right). \tag{12}$$

  for some function $G : \mathbb{R}^{n \times n} \to \mathbb{R}$.

- Most of the time the G function is a norm like the ℓ_1-norm or a power of a norm such as the ℓ_2 norm to the square.

Main general results for the statistical analysis of SDP estimators

- Function G should play the role of a *simple* description of the curvature of the excess risk function locally around Z^*; that is formalized in the next assumption.

Assumption 1

For all $Z \in \mathcal{C}$, *if*

$$\left\langle \mathbb{E}A, Z^* - Z \right\rangle \leq r_G^*(\Delta)$$

then

$$\left\langle \mathbb{E}A, Z^* - Z \right\rangle \geq G(Z^* - Z).$$

- Typical examples of curvature functions G will have the form $G(Z^* - Z) = \theta \left\| Z^* - Z \right\|^{\kappa}$ for some $\kappa \geq 1$, $\theta > 0$ and some norm $\left\| \cdot \right\|$.
  - In that case, the parameter κ was initially called the *margin parameter* Tsybakov 2003; Mammen and Tsybakov 1999.
- Even though the relation given in Assumption 1 has been typically referred to as a margin condition or Bernstein condition in the Learning Theory literature, we will rather call it a *local curvature assumption*, following Guédon and Vershynin 2016 and Chinot, Guillaume, and Matthieu 2018

Main general results for the statistical analysis of SDP estimators

- Using our curvature Assumption 1, we see that $r_G^*(\Delta)$ should be easier to compute than $r^*(\Delta)$ and $r^*(\Delta) \leq r_G^*(\Delta)$
  - This is because of the definition of $r_G^*(\Delta)$ and
$$\{Z \in \mathcal{C} : \left\langle \mathbb{E}A, Z^* - Z \right\rangle \leq r_G^*(\Delta)\} \subset \{Z \in \mathcal{C} : G(Z^* - Z) \leq r_G^*(\Delta)\}.$$
- Based on these remarks, we get the following

Corollary 2

We assume that the "local curvature" Assumption 1 holds for some $0 < \Delta < 1$. *With probability at least* $1 - \Delta$, *it holds true that*
$$r_G^*(\Delta) \geq \left\langle \mathbb{E}A, Z^* - \hat{Z} \right\rangle \geq G(Z^* - \hat{Z}).$$

Main general results for the statistical analysis of SDP estimators

Finally, when proving a "local curvature" property such as in Assumption 1 is too difficult we will replace Assumption 1 by

Assumption 2

For all $Z \in \mathcal{C}$, if $G(Z^* - Z) \leq r_G^*(\Delta)$ then $\langle \mathbb{E}A, Z^* - Z \rangle \geq G(Z^* - Z)$.

We can then obtain

Theorem 7

We assume that the constraint $\mathcal{C}$ is star-shaped in Z^* and that the "local curvature" Assumption 2 holds for some $0 < \Delta < 1$. We assume that the G function is continuous, $G(0) = 0$ and

$$G(\lambda(Z^* - Z)) \leq \lambda G(Z^* - Z)$$

for all $\lambda \in [0, 1]$ and $Z \in \mathcal{C}$. With probability at least $1 - \Delta$, it holds true that

$$G(Z^* - \hat{Z}) \leq r_G^*(\Delta).$$

Main general results for the statistical analysis of SDP estimators

As a consequence,

- Theorem 6, Corollary 2 and Theorem 7 are the three tools at our disposal to study the performance of SDP estimators depending on the deepness of understanding we have on the problem.
- The best approach is given by Theorem 6 when it is possible to compute efficiently the complexity fixed point $r^*(\Delta)$. If the latter approach is too complicated (likely because understanding the geometry of the local subset $\mathcal{C} \cap \{Z : \langle \mathbb{E}A, Z^* - Z \rangle \leq r\}, r > 0$ may be difficult) then one may resort to find a curvature function G of the excess risk locally around Z^*.
- In that case, both Corollary 2 and Theorem 7 may apply depending on the hardness to find a local curvature function G on an "excess risk neighborhood" (see Assumption 1) or a "G-neighborhood" (see Assumption 2).
- Finally, if no local approach can be handled (likely because describing the curvature of the excess risk in any neighborhood of Z^* or controlling the maximal oscillations of the empirical process $Z \to \langle \mathbb{E}A - A, Z^* - Z \rangle$ locally are too difficult) then one may resort ultimately to a global approach which follows from Theorem 6.

Main general results for the statistical analysis of SDP estimators

- Results like Theorem 6, Corollary 2 and Theorem 7 appeared in many papers on ERM in Learning Theory such as in Koltchinskii 2011; Bartlett and Mendelson 2006; Massart 2007; Lecué and Mendelson 2013.

- In all these results, typical loss functions such as the quadratic or logistic loss functions, were not linear ones, such as the one we are using here. Therefore, our problem is easier (and as a result, our three general results above at not so difficult to prove).

- What is much more complicated here than in other more classical problems in Learning Theory is the computation of the fixed point because
  - the stochastic processes $Z \to \left\langle A - \mathbb{E}A, Z - Z^* \right\rangle$ may be far from being a Gaussian process if the noise matrix $A - \mathbb{E}A$ is complicated and
  - the local sets
  $$\{Z \in \mathcal{C} : \left\langle \mathbb{E}A, Z^* - Z \right\rangle \le r\}$$
  or
  $$\{Z \in \mathcal{C} : G(Z^* - Z) \le r\}$$
  for $r > 0$ maybe very hard to describe in a simple way.

- Instrumental results are available in the literature to circumvent this kind of problems; see Fei and Chen 2019.

Revisiting two results from the community detection literature

- As we saw in the introduction, one challenging aspect of the community detection problem arises in the setting of sparse graphs.
- Many of the existing algorithms, which enjoy theoretical guarantees, do so in the relatively dense regime for the edge sampling probability, where the expected average degree is of the order $\Theta(\log n)$.
- The problem becomes challenging in very sparse graphs with bounded average degree.
  - To this end, Guédon and Vershynin 2016
    - proposed the semidefinite relaxation for the community detection problem (and more) we saw earlier
    - showed that it can recover a solution with any given relative accuracy even in the setting of very sparse graphs with average degree of order $O(1)$.

Revisiting two results from the community detection literature

- A subset of the existing literature for community detection and clustering relies on spectral methods, which consider the adjacency matrix associated to a graph, and employ its eigenvalues, and especially eigenvectors, in the analysis process or to propose efficient algorithms to solve the task at hand.

Revisiting two results from the community detection literature

- We now focus on the community detection problem on random graphs under the general stochastic block model.

- The results of Guédon and Vershynin 2016 and Fei and Chen 2019 can be approached using our Theorem 6.

- Thanks to this theorem, it is possible to simplify the proof of Fei and Chen 2019, by avoiding both the peeling argument and the use of the bound from Guédon and Vershynin 2016.

Revisiting two results from the community detection literature

We first recall the definition of the generalized stochastic block model (SBM). We consider a set of vertices $V = \{1, \cdots, n\}$, and assume it is partitioned into K communities $\mathcal{C}_1, \cdots, \mathcal{C}_K$ of arbitrary sizes $|\mathcal{C}_1| = l_1, \cdots, |\mathcal{C}_K| = l_K$.

Definition 2

For any pair of nodes $i, j \in V$, we denote by $i \sim j$ when i and j belong to the same community (i.e., there exists $k \in \{1, \ldots, K\}$) such that $i, j \in \mathcal{C}_k$), and we denote by $i \not\sim j$ if i and j do not belong to the same community.

- For each pair (i, j) of nodes from V, we draw an edge between i and j with a fixed probability p_{ij} independently from the other edges.
- We assume that there exist numbers p and q satisfying $0 < q < p < 1$, such that

$$\begin{cases} p_{ij} > p, \text{ if } i \sim j \text{ and } i \neq j, \\ p_{ij} = 1, \text{ if } i = j, \\ p_{ij} < q, \text{ otherwise.} \end{cases} \tag{13}$$

- We denote by $A = (A_{i,j})_{1 \leq i,j, \leq n}$ the observed symmetric adjacency matrix, such that, for all $1 \leq i \leq j \leq n$, A_{ij} is distributed according to a Bernoulli of parameter p_{ij}.

Revisiting two results from the community detection literature

- We will estimate its membership matrix $\bar{Z}$ via the following SDP estimator

  $$\hat{Z} \in \arg\max_{Z \in \mathcal{C}} \langle A, Z \rangle,$$

  where

  $$\mathcal{C} = \{Z \in \mathbb{R}^{n \times n}, Z \succeq 0, Z \geq 0, \mathrm{diag}(Z) \preceq I_n, \sum_{i,j=1}^{n} Z_{ij} \leq \lambda\}$$

  and

  $$\lambda = \sum_{i,j=1}^{n} \bar{Z}_{ij} = \sum_{k=1}^{K} |\mathcal{C}_k|^2$$

  (the number of nonzero elements in the membership matrix $\bar{Z}$).

- The motivation for this approach stems from the fact that the membership matrix $\bar{Z}$ is actually the oracle, i.e., $Z^* = \bar{Z}$, where

  $$Z^* \in \arg\max_{Z \in \mathcal{C}} \langle \mathbb{E}A, Z \rangle.$$

Revisiting two results from the community detection literature

- Guédon and Vershynin 2016 use the observation that, for all $r > 0$, it holds true that

$$\sup_{Z \in \mathcal{C} : \left\langle \mathbb{E}A, Z^* - Z \right\rangle \leq r} \left\langle A - \mathbb{E}A, Z - Z^* \right\rangle \overset{(a)}{\leq} \sup_{Z \in \mathcal{C}} \left\langle A - \mathbb{E}A, Z - Z^* \right\rangle \overset{(b)}{\leq} 2K_G \left\| A - \mathbb{E}A \right\|_{\mathrm{cut}} , \tag{14}$$

  where $\|\cdot\|_{\mathrm{cut}}$ is the cut-norm.

- The cut-norm $\|\cdot\|_{\mathrm{cut}}$ of a real matrix $A = (a_{ij})_{i \in R, j \in C}$ with a set of rows indexed by R and a set of columns indexed by C, is the maximum, over all $I \subset R$ and $J \subset C$, of the quantity $|\sum_{i \in I, j \in J} a_{ij}|$.
  - It is also the operator norm of A from ℓ_∞ to ℓ_1 and the "injective norm" in the orginal Grothendieck "résumé" Grothendieck 1956; Pisier 2012 and K_G is the Grothendieck constant (Grothendieck's inequality is used in (b), see Pisier 2012; Vershynin 2018).

Revisiting two results from the community detection literature

- The next step in the proof of Guédon and Vershynin 2016 is a high-probability upper bound on $\|A - \mathbb{E}A\|_{\text{cut}}$ which follows from Bernstein's inequality and a union bound since

  $$\|A - \mathbb{E}A\|_{\text{cut}} = \max_{x,y \in \{-1,1\}^n} \langle A - \mathbb{E}A, xy^\top \rangle,$$

  which implies that for all $t > 0$,

  $$\|A - \mathbb{E}A\|_{\text{cut}} \leq tn(n-1)/2$$

  with probability at least $1 - \exp\left(2n \log 2 - (n(n-1)t^2)/(16\bar{p} + 8t/3)\right)$ where

  $$\bar{p} \overset{def}{=} 2/[n(n-1)] \sum_{i<j} p_{ij}(1 - p_{ij}).$$

- The resulting upper bound on the fixed point obtained in Guédon and Vershynin 2016 is,

  $$r^*(\Delta) \leq (8/3)K_G(2n \log(2) + \log(1/\Delta)). \tag{15}$$

Revisiting two results from the community detection literature

- Finally, under the assumption of Theorem 1 in Guédon and Vershynin 2016 (i.e., for some some $\epsilon \in (0, 1)$, $n \geq 5.10^4/\epsilon^2$, $\max(p(1-p), q(1-q)) \geq 20/n$, $p = a/n > b/n = q$ and $(a-b)^2 \geq 2.10^4\epsilon^{-2}(a+b)$), for $\Delta = e^3 5^{-n}$ we obtain (using the general result in Theorem 6) with probability at least $1 - \Delta$, the bound

  $$\left\langle \mathbb{E}A, Z^* - \hat{Z} \right\rangle \leq r^*(\Delta) \leq \epsilon n^2 = \epsilon \left\| Z^* \right\|_2^2 ,$$

  which is the result from Theorem 1 in Guédon and Vershynin 2016

Revisiting two results from the community detection literature

- Finally, Guédon and Vershynin 2016 uses a (global) curvature property of the excess risk:

Lemma 2

For all $Z \in \mathcal{C}$, $\langle \mathbb{E}A, Z^* - Z \rangle \geq [(p-q)/2] \|Z^* - Z\|_1$.

Therefore, a (global– that is for all $Z \in \mathcal{C}$) curvature assumption holds for a G function which is here the $\ell_1^{n\times n}$ norm, a margin parameter $\kappa = 1$ and $\theta = (p-q)/2$ for the community detection problem.

- However, this curvature property is
  - not used to compute a "better" fixed point parameter
  - but only to obtain a $\ell_1^{n\times n}$ estimation bound since

$$\left\|\hat{Z} - Z^*\right\|_1 \leq \left(\frac{2}{p-q}\right) \langle \mathbb{E}A, Z^* - \hat{Z} \rangle \leq \frac{16K_G(2n\log(2) + \log(1/\Delta))}{3(p-q)}.$$

- The latter bound together with the sin-Theta theorem allow the authors of Guédon and Vershynin 2016 to obtain estimation bound for the community membership vector x^*.

Revisiting two results from the community detection literature

- The approach from Fei and Chen 2019 improves upon the one in Guédon and Vershynin 2016 because it uses a localization argument
- Indeed, the authors from Fei and Chen 2019 obtain high-probability upper bound on the quantity
$$\sup_{Z\in\mathcal{C}:\|Z^*-Z\|_1\leq r}\left\langle A-\mathbb{E}A, Z-Z^*\right\rangle$$
depending on r.
- This yields to exact reconstruction result in the "dense" case and exponentially decaying rates of convergence in the "sparse" case.
    - This is a typical example where the localization argument shows its advantage upon the global approach.
- However, the argument from Fei and Chen 2019 also uses an unnecessary peeling argument together with an unnecessary a priori upper bound on $\left\|\hat{Z}-Z^*\right\|_1$
- It appears that this peeling argument and this a priori upper bound on $\left\|\hat{Z}-Z^*\right\|_1$ can be avoided thanks to our approach from Theorem 6.
- This improves the probability estimate and simplifies the proofs (since the result from Guédon and Vershynin 2016 is not required anymore neither is the peeling argument).

OTHER EXAMPLES

We now study three other problems, namely

- signed clustering,
- angular synchronization, and
- MAX-CUT.

Other examples

A Signed Stochastic Block Model (SSBM)

- We focus on the problem of clustering a K-weakly balanced graphs
- A signed graph is K-weakly balanced if and only if all the edges are positive, or the nodes can be partitioned into $K \in \mathbb{N}$ disjoint sets such that positive edges exist only within clusters, and negative edges are only present across clusters.
- We consider a signed stochastic block model (SSBM) similar to the one introduced in Cucuringu et al. n.d., where we are given
  - a graph G with n nodes $\{1, \ldots, n\}$ which are divided into K communities, $\{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$, such that, in the noiseless setting,
    - edges within each community are positive and
    - edges between communities are negative.

Other examples

A Signed Stochastic Block Model (SSBM)

- The only information available to the user is given by a $n \times n$ sparse censored signed adjacency matrix A constructed as follows:
  - A is symmetric, with $A_{ii} = 1$ for all $i = 1, \ldots, n$, and for all $1 \leq i < j \leq n$, $A_{ij} = s_{ij}(2B_{ij} - 1)$ where

    $$\mathrm{B}_{ij} \sim \begin{cases} \text{Bern}(p) \text{ if } i \sim j \\ \\ \text{Bern}(q) \text{ if } i \not\sim j \end{cases} \quad \text{and} \quad s_{ij} \sim \text{Bern}(\delta),$$

    for some $0 \leq q < 1/2 < p \leq 1$ and $\delta \in (0, 1)$. Moreover, all the variables B_{ij}, s_{ij} for $1 \leq i < j \leq n$ are assumed independent.
- Our aim is to recover the community membership matrix (or cluster matrix) $\bar{Z} = (\bar{Z}_{ij})_{i,j \leq n}$, with

  $$\begin{cases} \bar{Z}_{ij} = 1 \text{ when } i \sim j \text{ and} \\ \\ \bar{Z}_{ij} = 0 \text{ when } i \not\sim j. \end{cases}$$

  using only the observed censored adjacency matrix A.

Other examples

A Signed Stochastic Block Model (SSBM)

- Our approach is similar in nature to the one used by spectral methods in community detection.
- We first observe that for $\alpha := \delta(p+q-1)$ and $J = (1)_{n\times n}$ we have $\bar{Z} = Z^*$ where

$$Z^* \in \arg\max_{Z\in\mathcal{C}} \langle \mathbb{E}A - \alpha J, Z \rangle \tag{16}$$

  and $\mathcal{C} = \{Z \in \mathbb{R}^{n\times n} : Z \succeq 0, Z_{ij} \in [0,1], Z_{ii} = 1, i = 1, \ldots, n\}$.

Other examples

A Signed Stochastic Block Model (SSBM)

- Since we do not know $\mathbb{E}A$ and α, we should estimate both of them. We will estimate $\mathbb{E}A$ with A but, for simplicity, we will assume that α is known. The resulting estimator of the cluster matrix $\bar{Z}$ is
$$\hat{Z} \in \operatorname*{arg\,max}_{Z \in \mathcal{C}} \langle A - \alpha J, Z \rangle \tag{17}$$
which is indeed a SDP estimator and therefore Theorem 6 (or Corollary 2 and Theorem 7) may be used to obtain statistical bounds for the estimation of Z^* from (16) by $\hat{Z}$.

Other examples

A Signed Stochastic Block Model (SSBM)

- Our main result concerns the reconstitution of the K communities from the observation of the matrix A.
  - In order to avoid solutions with some communities of degenerated size (too small or too large) we consider the following assumption.

Assumption 3

Up to constants, the elements of the partition $\mathcal{C}_1 \sqcup \cdots \sqcup \mathcal{C}_K$ of $\{1, \ldots, n\}$ are of same size: there are absolute constant $c_0, c_1 > 0$ such that for any $k \in [K]$, $n/(c_1 K) \leq |\mathcal{C}_k| = l_k \leq c_0 n/K$.

Other examples

A Signed Stochastic Block Model (SSBM)

We are now ready to state the main result on the estimation of the cluster matrix Z^* from (16) by the SDP estimator $\hat{Z}$ from (17).

Theorem 8

There is an absolute positive constant c_0 such that the following holds. Grant Assumption 3. Assume that

$$n\nu\delta \geq \log n, \tag{18}$$

$$sn \geq c_0 K^2 \nu \tag{19}$$

$$\textit{and} \qquad \frac{K\log(2eKn)}{n} \leq \max\left(\frac{\theta^2}{\rho}, \frac{9\rho}{32}\right). \tag{20}$$

Then, with probability at least $1 - \exp(-\delta\nu n) - 3/(2eKn)$, exact recovery holds true, i.e., $\hat{Z} = Z^*$.

We recall the constants defined above :

$$s := \delta(p-q)^2,\ \theta := \delta(p-q),\ \rho := \delta\max\{1-\delta(2p-1)^2, 1-\delta(2q-1)^2\},\ \nu := \max\{2p-1, 1-2q\}.$$

Other examples

A Signed Stochastic Block Model (SSBM)

- The last condition (20) basically requires that the number of clusters K is at most $n/\log n$.
- If this condition is dropped out, then we do not have anymore exact reconstruction but only a certified exponential rate of convergence:
  - there exists a universal constant C_2 such that, with probability at least $1 - \exp(-\delta\nu n) - 3/(2eKn)$,

$$\left\| Z^* - \hat{Z} \right\|_1 \leq \frac{2\mathrm{e}n^2}{c_1\theta} \exp\left(-\frac{sn}{C_2 K} \right). \tag{21}$$

Other examples

A Signed Stochastic Block Model (SSBM)

- This shows that there is a phase transition in the dense case:
  - exact reconstruction is possible when $K \lesssim n/\log n$ and,
  - otherwise, when $K \gtrsim n/\log n$ we only have a control of the estimation error with an exponential convergence rate.
- We then get results of the same nature as in Fei and Chen 2019, or in the more recent paper M. Xu et al. 2020.
  - In those two articles, the authors show the existence of a phase transition, with exact recovery in the regime $K \lesssim n/log(n)$, and exponential rate with exponent $\simeq -sn/K$ otherwise, where s is some measurement of the signal/noise ratio of the problem.
  - Note that the estimation bound is given with respect to the $l_1^{n\times n}$ norm. This is not a surprise since it is the behaviour of the excess risk over $\mathcal{C}$ and Z^*.

Other examples

Application to angular group synchronization

- We now introduce the group synchronization problem as well as a stochastic model for this problem. We consider a SDP relaxation of the original problem (which is exact) and construct the associated SDP estimator such as in (10).
- The angular synchronization problem consists of estimating n unknown angles $\theta_1, \cdots, \theta_n$ (up to a global shift angle) given a noisy subset of their pairwise offsets $(\theta_i - \theta_j)[2\pi]$, where $[2\pi]$ is the modulo 2π operation.
- The pairwise measurements can be realized as the edge set of a graph G, typically modeled as an Erdös-Renyi random graph Singer 2011.

Other examples

Application to angular group synchronization

- The aim of this section is to show that the angular synchronization problem can be analyzed using our methodology.

- In order to keep the presentation as simple as possible, we assume that all pairwise offsets are observed up to some Gaussian noise: we are given $\delta_{ij} = (\theta_i - \theta_j + \sigma g_{ij})[2\pi]$ for all $1 \le i < j \le n$ where $(g_{ij} : 1 \le i < j \le n)$ are $n(n-1)/2$ i.i.d. standard Gaussian variables and $\sigma > 0$ is the noise variance.

- We may rewrite the problem as follows: we observe a $n \times n$ complex matrix A defined by

$$A = S \circ [x^*(\overline{x^*})^\top] \text{ where } S = (S_{ij})_{n\times n}, S_{ij} = \begin{cases} e^{\iota\sigma g_{ij}} & \text{if } i < j \\ 1 & \text{if } i = j \\ e^{-\iota\sigma gij} & \text{if } i > j \end{cases}, \tag{22}$$

  ι denotes the imaginary number such that $\iota^2 = -1$, $x^* = (x_i^*)_{i=1}^n \in \mathbb{C}^n$, $x_i^* = e^{\iota\theta_i}, i = 1, \dots, n$, $\bar{x}$ denotes the conjugate vector of x and $S \circ [x^*(\overline{x^*})^\top]$ is the element-wise product $(S_{ij}x_i\bar{x}_j)_{n\times n}$.

- In particular, S is a Hermitian matrix (i.e. $\bar{S}^\top = S$) and $\mathbb{E}S_{ij} = \exp(-\sigma^2/2)$ for $i \neq j$ and $\mathbb{E}S_{ii} = 1$ if $i = j$.

- We want to estimate $(\theta_1, \dots, \theta_n)$ (up to a global shift) from the matrix of data A.

Other examples

Application to angular group synchronization

- Estimating $(\theta_i)_{i=1}^n$ up to global angle shift is equivalent to estimating the vector $x^* = (e^{\iota\theta_i})_{i=1}^n$.
- The latter is, up to a global rotation of its coordinates, the unique solution of the following maximization problem

$$\operatorname*{arg\,max}_{x\in\mathbb{C}^n:|x_i|=1} \left\{\bar{x}^\top \; \mathbb{E}A \; x\right\} = \{(e^{\iota(\theta_i+\theta_0)})_{i=1}^n : \theta_0 \in [0, 2\pi)\}. \tag{23}$$

- Let us now rewrite (23) as a SDP problem.
  - For all $x \in \mathbb{C}^n$, we have $\bar{x}^\top \mathbb{E}Ax = \mathrm{tr}(\mathbb{E}AX) = \langle \mathbb{E}A, X\rangle$ where $X = x\bar{x}^\top$ and $\{Z \in \mathbb{C}^{n\times n} : Z = x\bar{x}^T, |x_i| = 1\} = \{Z \in \mathbb{H}_n : Z \succeq 0, \mathrm{diag}(Z) = \mathbf{1}_n, \mathrm{rank}(Z) = 1\}$ where $\mathbb{H}_n$ is the set of all $n \times n$ Hermitian matrices and $\mathbf{1}_n \in \mathbb{C}^n$ is the vector with all coordinates equal to 1.

  - It is therefore straightforward to construct a SDP relaxation of (23) by dropping the rank constraint.

  - It appears that this relaxation is exact since, for $\mathcal{C} = \{Z \in \mathbb{H}_n : Z \succeq 0, \mathrm{diag}(Z) = \mathbf{1}_n\}$,

$$\operatorname*{arg\,max}_{Z\in\mathcal{C}} \langle \mathbb{E}A, Z\rangle = \{Z^*\}, \tag{24}$$

where $Z^* = x^*(\overline{x^*})^\top$.

Other examples

Application to angular group synchronization

- Finally, as we only observe A, we consider the following SDP estimator of Z^*

$$\hat{Z} \in \arg\max_{Z \in \mathcal{C}} \langle A, Z \rangle. \tag{25}$$

Other examples

Application to angular group synchronization

- Our main result concerns the estimation of the matrix of offsets $Z^* = x^*(\overline{x^*})^\top$ from the observation of the matrix A.

- This result is then used to estimate (up to a global phase shift) the angular vector $x^* = (e^{-\iota\theta_i})_{i=1}^n$.

- Our first result follows from Corollary 2.

Theorem 9

Let $0 < \epsilon < 1$*. If* $\sigma \leq \sqrt{\log(\epsilon n^4)}$ *then, with probability at least* $1 - \exp(-\epsilon\sigma^4 n(n-1)/2)$*, it holds true that*

$$(e^{-\sigma^2/2}/2)\,\|Z^* - Z\|_2^2 \leq \left\langle \mathbb{E}A, Z^* - Z \right\rangle \leq (128/6)\sqrt{\epsilon}\sigma^4 n(n-1). \tag{26}$$

Other examples

Application to angular group synchronization

Once we have an estimator $\hat{Z}$ for the oracle Z^*, we can extract an estimator $\hat{x}$ for the vector of phases x^* by considering a top eigenvector (i.e. an eigenvector associated with the largest eigenvalue) of $\hat{Z}$. It is then possible to quantify the estimation properties of x^* by $\hat{x}$ using a sin-Theta theorem and Theorem 9.

Corollary 3

Let $\hat{x}$ be a top eigenvector of $\hat{Z}$ with Euclidean norm $\|\hat{x}\|_2 = \sqrt{n}$. Let $0 < \epsilon < 1$ and assume that $\sigma \leq \sqrt{\log(\epsilon n^4)}$. We have the existence of a universal constant $c_0 > 0$ (which is the constant in the Davis-Kahan theorem for Hermitian matrices) such that, with probability at least $1 - \exp(-\epsilon\sigma^4 n(n-1)/2)$, it holds true that

$$\min_{z \in \mathbb{C}:|z|=1} \|\hat{x} - zx^*\|_2 \leq 8c_0\sqrt{2/3}\epsilon^{1/4}e^{\sigma^2/4}\sigma^2\sqrt{n}. \quad (27)$$

Other examples

Application to the MAX-CUT problem

- Let $A^0 \in \{0,1\}^{n\times n}$ be the adjacency (symmetric) matrix of an undirected graph $G = (V, E^0)$, where $V := \{1, \ldots, n\}$ is the set of the vertices and the set of edges is $E^0 := E \cup E^\top \cup \{(i,i) : A^0_{ii} = 1\}$ where $E := \{(i,j) \in V^2 \; : \; i < j$ and $A^0_{ij} = 1\}$ and $E^\top = \{(j,i) : (i,j) \in E\}$.

- We assume that G has no self loop so that $A^0_{ii} = 0$ for all $i \in V$. A *cut* of G is any subset S of vertices in V.

- For a cut $S \subset V$, we define its weight by $\mathrm{cut}(G,S) := (1/2)\sum_{(i,j)\in S\times\bar{S}} A^0_{ij}$, that is the number of edges in E between S and its complement $\bar{S} = V \backslash S$.

- The MAX-CUT problem is to find the cut with maximal weight:

$$S^* \in \operatorname*{argmax}_{S\subset V} \mathrm{cut}(G,S). \tag{28}$$

Other examples

Application to the MAX-CUT problem

- The MAX-CUT problem is a NP-complete problem but Goemans and Williamson 1995 constructed a 0.878 approximating solution via a SDP relaxation. Indeed, one can write the MAX-CUT problem in the following way.
- For a cut $S \subset V$, we define the membership vector $x \in \{-1, 1\}^n$ associated with S by setting $x_i := 1$ if $i \in S$ and $x_i = -1$ if $i \notin S$ for all $i \in V$. We have $\mathrm{cut}(G, S) = (1/4) \sum_{i,j=1}^n A^0_{ij}(1 - x_i x_j) := \mathrm{cut}(G, x)$ and so solving (28) is equivalent to solving

$$x^* \in \underset{x \in \{-1,1\}^n}{\mathrm{argmax}}\ \mathrm{cut}(G, x). \tag{29}$$

- Since $(x_i x_j)_{i,j} = xx^\top$, the latter problem is also equivalent to solving

$$\max \left(\frac{1}{4} \sum_{i,j=1}^n A^0_{ij}(1 - Z_{ij}) : \mathrm{rank}(Z) = 1, Z \succeq 0, Z_{ii} = 1 \right) \tag{30}$$

  which admits a SDP relaxation by removing the rank 1 constraint.
- This yields the following SDP relaxation problem of MAX-CUT from Goemans and Williamson 1995:

$$Z^* \in \underset{Z \in \mathcal{C}}{\mathrm{argmin}} \langle A^0, Z \rangle \tag{31}$$

  where $\mathcal{C} := \{Z \in \mathbb{R}^{n \times n} : Z \succeq 0, Z_{ii} = 1, \forall i = 1, \ldots, n\}$.

Other examples

Application to the MAX-CUT problem

- Unlike the other examples from the previous sections, the SDP relaxation in (31) is not exact, except for bipartite graphs; see Khot and Naor 2009; Gärtner and Matoušek 2012 for more details.

- Nevertheless, thanks to the approximation result from Goemans and Williamson 1995, we can use our methodology to estimate Z^* and then deduce an approximate optimal cut.

- The MAX-CUT problem is therefore a good setup for us to test our methodology in a context where the SDP relaxation is not exact, but still widely used in practice.

Other examples

Application to the MAX-CUT problem

- Thus the type of question we want to answer here is: what can we say in a setup where only partial or noisy information is available on $\mathbb{E}[A]$, and when the SDP relaxation associated with $\mathbb{E}[A]$ is also not exact?
- This differs from the previous setup where exactness of the SDP relaxation holds, and this interesting peculiarity is one of the reasons why we have chosen to present this problem here.
- Our motivation stems from the observation that, in many situations, the adjacency matrix A^0 is only partially observed, but nevertheless, it might be interesting to find an approximating solution to the MAX-CUT.

Other examples

Application to the MAX-CUT problem

- Let us then introduce a stochastic model for the partial information available on $\mathbb{E}[A]$, the adjacency matrix here.

- We observe $A = S \circ A^0 = (s_{ij}A^0_{ij})_{1\leq i,j\leq n}$ a "masked" version of A^0, where $S \in \mathbb{R}^{n\times n}$ is symmetric with upper triangular matrix filled with i.i.d. Bernoulli entries: for all $i, j \in V$ such that $i \leq j$, $S_{ij} = S_{ji} = s_{ij}$ where $(s_{ij})_{i\leq j}$ is a family of i.i.d. Bernoulli random variables with parameter $p \in (1/2, 1)$.

- Let $B := -(1/p)A$ so that $\mathbb{E}[B] = -A^0$.

- We can write Z^* as an oracle since $Z^* \in \arg\max_{Z\in\mathcal{C}} \langle \mathbb{E}B, Z\rangle$ and so we estimate Z^* via the SDP estimator $\hat{Z} \in \arg\max_{Z\in\mathcal{C}} \langle B, Z\rangle$.

Other examples

Application to the MAX-CUT problem

- Our first aim is to quantify the cost we pay by using $\hat{Z}$ instead of Z^* in our final choice of cut.
- It appears that the fixed point used in Theorem 6 may be used to quantify this loss

$$r^*(\Delta) = \inf\left(r > 0 : \mathbb{P}\left[\sup_{Z \in \mathcal{C} : \left\langle \mathbb{E}B, Z^* - Z \right\rangle \leq r} \left\langle B - \mathbb{E}B, Z - Z^* \right\rangle \leq (1/2)r\right] \geq 1 - \Delta\right). \qquad (32)$$

Other examples

Application to the MAX-CUT problem

- We denote the optimal values of the MAX-CUT problem associated with the graph G and its SDP relaxation by

$$\mathrm{SDP}(G) := (1/4)\langle A^0, J - Z^* \rangle = \max_{Z \in \mathcal{C}} \frac{1}{4} \sum_{i,j} A^0_{i,j}(1 - Z_{ij}) \text{ and } \mathrm{MAXCUT}(G) := \mathrm{cut}(G, S^*)$$

where S^* is a solution of (28) and $J = (1)_{n \times n}$.

Other examples

Application to the MAX-CUT problem

- Our first result is to show how the 0.878 approximating result from Goemans and Williamson 1995 is downgraded by the incomplete information we have on the graph (we only partially observed the adjacency matrix A^0 via the masked matrix A).

Theorem 10

For all $0 < \Delta < 1$*. With probability at least* $1 - \Delta$ *(with respect to the masked* S*),*

$$\mathrm{SDP}(G) \geq \mathbb{E}\left[\mathrm{cut}(G, \hat{x}) | \hat{Z}\right] \geq 0.878\mathrm{SDP}(G) - \frac{0.878 r^*(\Delta)}{4}.$$

- To make the notation more precise, $\hat{x}$ is the sign vector of $\hat{G}$ which is a centered Gaussian variable with covariance $\hat{Z}$.
- In that context, $\mathbb{E}\left[\mathrm{cut}(G, \hat{x}) | \hat{Z}\right]$ is the conditional expectation according to $\hat{G}$ for a fixed $\hat{Z}$.
- Moreover, the probability “at least $1 - \Delta$" that we obtain is w.r.t. the random masks, that is to the randomness in A.

Other examples

Application to the MAX-CUT problem

- Let us now frame Theorem 10 into the following perspective. If we had known the entire adjacency matrix (which is the case when $p = 1$), then we could have used Z^* instead of $\hat{Z}$. In that case, for $x^\star$ the sign vector of $G^\star \sim \mathcal{N}(0, Z^*)$, we know from Goemans and Williamson 1995 that
$$\mathrm{SDP}(G) \geq \mathbb{E}\left[\mathrm{cut}(G, x^\star)\right] \geq 0.878\mathrm{SDP}(G). \tag{33}$$

- Therefore, Theorem 10 characterizes the price we pay for not observing the entire adjacency matrix A^0, but only a masked version A of it.

- It is an interesting output of Theorem 10 to observe that the fixed point $r^*(\Delta)$ measures, in a quantitative way, this loss.

- If we were able to identify scenarios of p and E for which $r^*(\Delta) = 0$, that would prove that there is no loss for partially observing A^0 in the MAX-CUT problem.

- The approach we use to control $r^*(\Delta)$ is the global one, which does not allow for exact reconstruction (that is, to show that $r^*(\Delta) = 0$).

Other examples

Application to the MAX-CUT problem

Let us now turn to an estimation result of Z^* by $\hat{Z}$ via an upper bound on $r^*(\Delta)$.

Theorem 11

With probability at least $1 - 4^{-n}$*:*

$$\left\langle \mathbb{E}B, Z^* - \hat{Z} \right\rangle \leq r^*(4^{-n}) \leq 2n\sqrt{\frac{(2\log 4)(1-p)(n-1)}{p}} + \frac{8n\log 4}{3}.$$

In particular, it follows from the approximation result from Theorem 10 and the high-probability upper bound on $r^*(\Delta)$ from Theorem 11 that, with probability at least $1 - 4^{-n}$

$$\mathbb{E}\left[\mathrm{cut}(G,\hat{x})|\hat{Z}\right] \geq 0.878\mathrm{SDP}(G) - \frac{0.878}{4}\left(2n\sqrt{\frac{(2\log 4)(1-p)(n-1)}{p}} + \frac{8n\log 4}{3}\right). \tag{34}$$

Other examples

Application to the MAX-CUT problem

- This result is non-trivial only when the right-hand side term is strictly larger than $0.5 \cdot \mathrm{SDP}(G)$, which is the performance of a random cut.
- As a consequence, (34) shows that one can still do better than randomness even in an incomplete information setup for the MAX-CUT problem when p, n and $\mathrm{SDP}(G)$ are such that

$$0.378\mathrm{SDP}(G) > \frac{0.878}{4}\left(2n\sqrt{\frac{(2\log 4)(1-p)(n-1)}{p}} + \frac{8n\log 4}{3}\right).$$

- For instance, when p is like a constant, it requires $\mathrm{SDP}(G)$ to be larger than $c_0 n^{3/2}$ (for some absolute constant c_0) and when $p = 1 - 1/n$, it requires $\mathrm{SDP}(G)$ to be at least $c_0 n$ (for some absolute constant c_0).

Other examples

Application to the MAX-CUT problem

Remark 1

- *To get exact recovery, that is $r^*(\Delta) = 0$, in the MAX-CUT problem (which shows that there is no loss for the* MAX-CUT *problem by observing only a masked version of the adjacency matrix), we have to develop a local approach, as for the*
  1. Signed Clustering and
  2. the Group Synchronization problems.
- *To that end, we would need to solve the following two problems:*
  1. Find a curvature for the objective function $Z \to \langle \mathbb{E}B, Z^* - Z \rangle$ and
  2. Study the oscillations of the empirical process $Z \to \langle \mathbb{E}B - B, Z^* - Z \rangle$.

We leave those two difficult problems for future research.

Part IV

Experiments

Numerical experiments

Signed Clustering

- Consider the following experimental setup.
  - We generate synthetic networks following the signed stochastic block model (SSBM) with $K = 5$ communities.
  - To quantify the effectiveness of the SDP relaxation, compare the accuracy of a suite of algorithms from signed clustering literature,
    - *before* the SDP relaxation (i.e., when we perform these algorithms directly on A) and
    - *after* the SDP relaxation (i.e., when we perform the very same algorithms on $\hat{Z}$).
  - Overall, (**??**) essentially counts the fraction of intra-cluster and inter-cluster edge violations, with respect to the full ground truth matrix.

Numerical experiments

Signed Clustering

- In terms of the signed clustering algorithms compared, we consider the following algorithms from the literature.
  - One straightforward approach is to simply rely on the spectrum of the observed adjacency matrix A. Kunegis et al. 2010 proposed spectral tools for clustering, link prediction, and visualization of signed graphs, by solving a 2-way "signed" ratio-cut problem based on the combinatorial Signed Laplacian Hou 2005 $\bar{L} = \bar{D} - A$, where $\bar{D}$ is a diagonal matrix with $\bar{D}_{ii} = \sum_{i=1}^{n} |A_{ij}|$.
  - The same authors proposed signed extensions for the case of the random-walk Laplacian $\bar{L}_{\text{rw}} = I - \bar{D}^{-1}A$, and the symmetric graph Laplacian $\bar{L}_{\text{sym}} = I - \bar{D}^{-1/2}A\bar{D}^{-1/2}$, the latter of which is particularly suitable for skewed degree distributions.
  - Finally, the last algorithm we considered is BNC of Chiang, Whang, and Dhillon 2012, who introduced a formulation based on the *Balanced Normalized Cut* objective

$$\min_{\{x_1,\ldots,x_K\}\in\mathcal{I}} \left(\sum_{c=1}^{K} \frac{x_c^T(D^+ - A)x_c}{x_c^T\bar{D}x_c} \right), \tag{35}$$

    which, in light of the decomposition $D^+ - A = D^+ - (A^+ - A^-) = D^+ - A^+ + A^- = L^+ + A^-$, is effectively minimizing the number of violations in the clustering procedure.

Numerical experiments

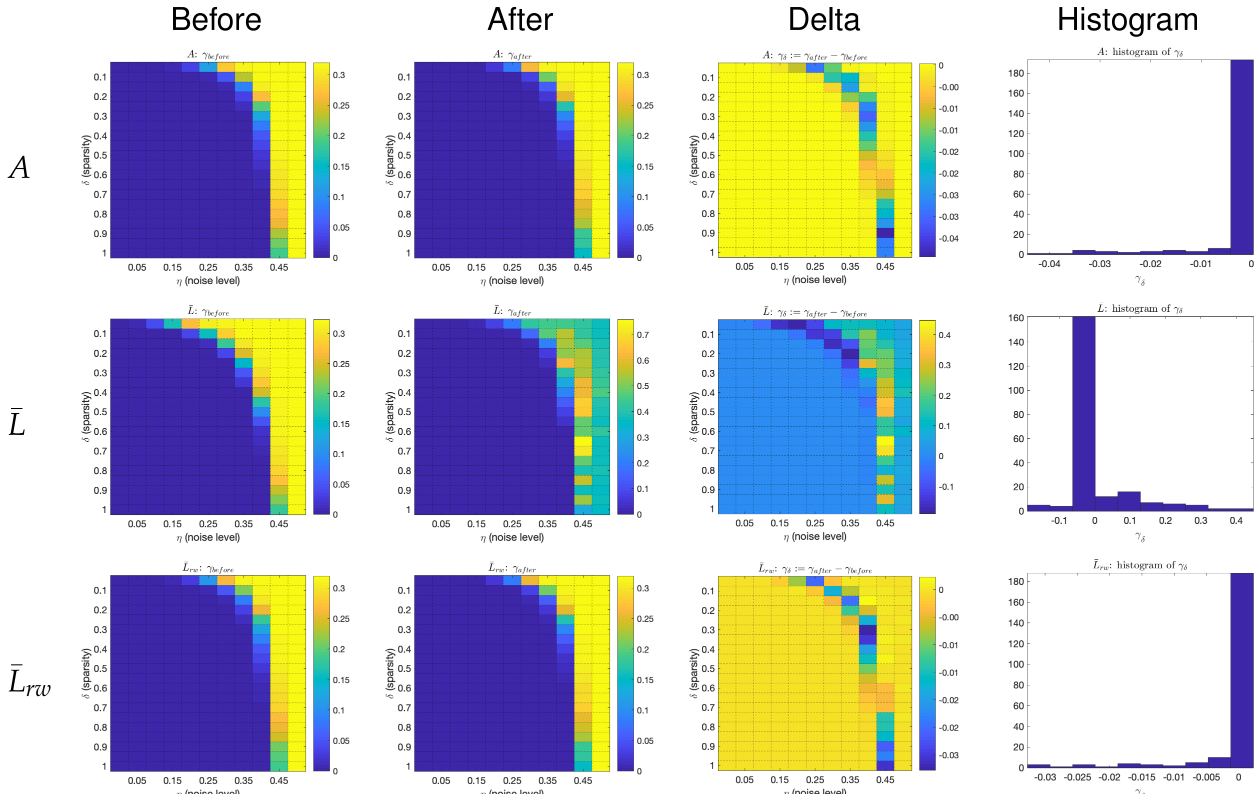
Signed Clustering

- In our experiments, we first compute the error rate γ_{before} of all algorithms on the original SSBM graph, and then we repeat the procedure but with the input to all signed clustering algorithms being given by the output of the SDP relaxation, and denote the resulting recovery error by γ_{after}.
- The third column of the next Figure shows the difference in errors $\gamma_\delta = \gamma_{before} - \gamma_{after}$ between the first and second columns, while the fourth column contains a histogram of the error differences γ_δ.

Numerical experiments

Signed Clustering

[!htp]



NUMERICAL EXPERIMENTS

SIGNED CLUSTERING

- These experiments altogether illustrate the fact that the SDP relaxation does improve the performance of all signed clustering algorithms, except $\bar{L}$, and could effectively be used as a denoising pre-processing step.
    - One potential reason why the SDP pre-processing step does not improve on the accuracy of $\bar{L}$ could stem from the fact that $\bar{L}$ has a good performance to begin with on examples where the clusters have equal sizes and the degree distribution is homogeneous.
    - It would be interesting to further compare the results in settings with skewed degree distributions, such as the classical Barabási-Albert model Albert and Barabási 2002.

NUMERICAL EXPERIMENTS
MAX-CUT

- For the MAX-CUT problem, we consider two sets of numerical experiments.
- First, we consider a version of the stochastic block model which essentially perturbs a complete bipartite graph
$$B = \begin{vmatrix} 0_{n_1 \times n_1} & 1_{n_1 \times n_2} \\ 1_{n_2 \times n_1} & 0_{n_2 \times n_2} \end{vmatrix}, \tag{36}$$
where $1_{n_1 \times n_2}$ (respectively, $0_{n_1 \times n_2}$) denotes an $n_1 \times n_2$ matrix of all ones, respectively, all zeros.
- In our experiments, we set $n_1 = n_2 = \frac{n}{2}$, and fix $n = 500$. We perturb B by deleting edges across the two partitions, and inserting edges within each partition.
- More specifically, we generated the *full* adjacency matrix A^0 from B by adding edges independently with probability η within each partition (i.e., along the diagonal blocks in (36)).

Numerical experiments
MAX-CUT

- Finally, we denote by A the masked version we observe, $A = A^0 \circ S$, where S denotes the adjacency matrix of an Erdős-Rényi(n, δ) graph. The graph shown in Figure 13 is an instance of the above generative model.

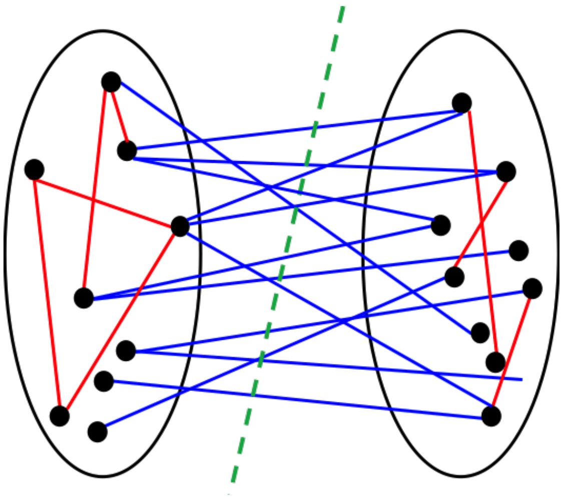


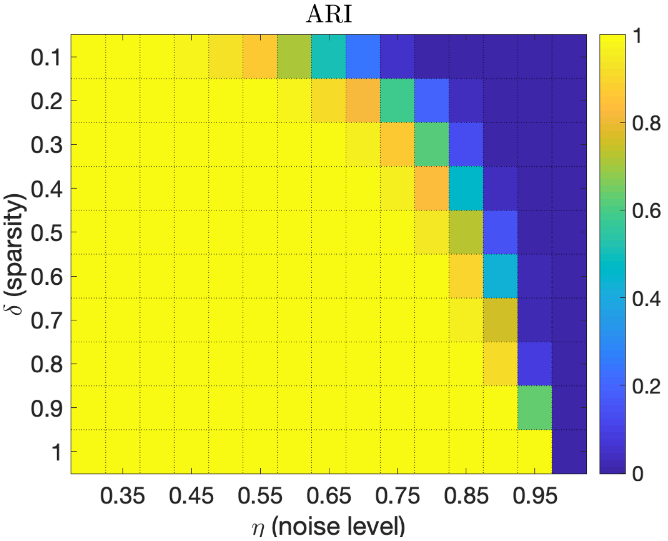
Figure. Illustration of MAX-CUT in the setting of a perturbation of a complete bipartite graph.

- Note that, for small values of η, we expect the maximum cut to occur across the initial partition $\mathcal{P}_B$ in the clean bipartite graph B, which we aim to recover as we sparsify the observed graph A.

Numerical experiments

MAX-CUT

- As expected, for a fix level of noise η, we are able to recover the hypothetically optimal MAX-CUT, for suitable levels of the sparsity parameter.

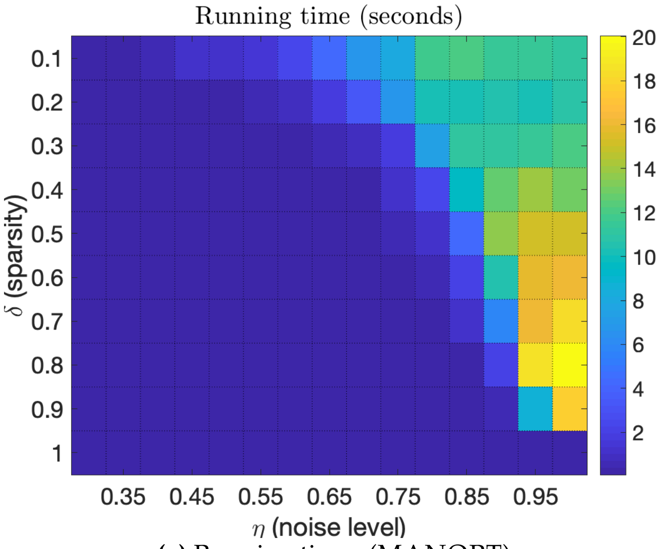


(a) Adjusted Rand Index.

Figure. Numerical results for MAX-CUT on a perturbed complete bipartite graph, as we vary the noise level η and the sampling sparsity δ. Results are averaged over 20 runs.

Numerical experiments
MAX-CUT

- The heatmap shows the computational running time, as we vary the two parameters, showing that the Manopt solver takes the longest to solve dense noisy problems, as one would expect.

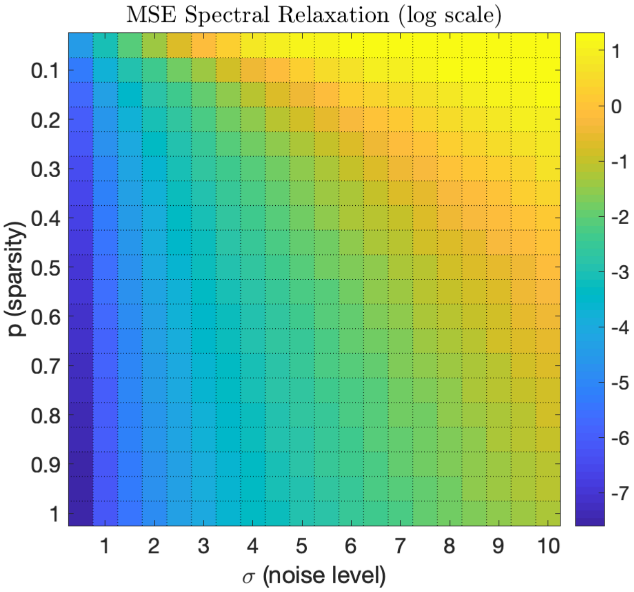


(a) Running times (MANOPT).

Figure. Numerical results for MAX-CUT on a perturbed complete bipartite graph, as we vary the noise level η and the sampling sparsity δ. Results are averaged over 20 runs.

Numerical experiments

Angular Synchronization

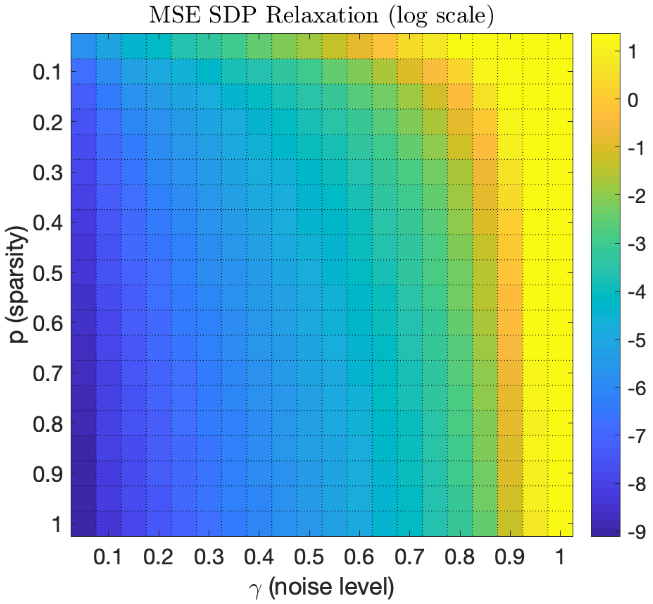


(a) Spectral relaxation.

Figure. Recovery rates (MSE (**??**) - the lower the better) for angular synchronization with $n = 500$, under the Gaussian noise model, as we vary the noise level σ and the sparsity p of the measurement graph. Results are averaged over 20 runs.

Numerical experiments

Angular Synchronization



(a) SDP relaxation (solved via MANOPT).

Figure. Recovery rates (MSE (**??**) - the lower the better) for angular synchronization with $n = 500$, under the Outlier noise model, as we vary the noise level γ and the sparsity p of the measurement graph. Results are averaged over 20 runs.

References I

- Abbe, Emmanuel, Afonso S Bandeira, and Georgina Hall (2015). “Exact recovery in the stochastic block model”. In: *IEEE Transactions on Information Theory* 62.1, pp. 471–487.
- Albert, Réka and Albert-László Barabási (2002). “Statistical mechanics of complex networks”. In: *Reviews of modern physics* 74.1, p. 47.
- Ames, Brendan PW (2014). “Guaranteed clustering and biclustering via semidefinite programming”. In: *Mathematical Programming* 147.1-2, pp. 429–465.
- Bartlett, Peter L. and Shahar Mendelson (2006). “Empirical minimization”. In: *Probab. Theory Related Fields* 135.3, pp. 311–334. ISSN: 0178-8051.
- Boyd, Stephen, Laurent El Ghaoui, et al. (1994). *Linear matrix inequalities in system and control theory*. Vol. 15. Siam.
- Boyd, Stephen and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.
- Chiang, Kai-Yang, Joyce Whang, and Inderjit S. Dhillon (Oct. 2012). “Scalable Clustering of Signed Networks using Balance Normalized Cut”. In: *ACM Conference on Information and Knowledge Management (CIKM)*.
- Chinot, Geoffrey, Lecué Guillaume, and Lerasle Matthieu (2018). “Statistical learning with Lipschitz and convex loss functions”. In: *arXiv preprint arXiv:1810.01090*.
- Chrétien, Stéphane and Franck Corset (2009). “Using the eigenvalue relaxation for binary least-squares estimation problems”. In: *Signal Processing* 89.11, pp. 2079–2091.

References II

- Chrétien, Stéphane, Clément Dombry, and Adrien Faivre (to appear). "A semi-definite programming approach to low dimensional embedding for unsupervised clustering". In: *Frontiers in Applied Mathematics and Statistics*.
- Cucuringu, M. et al. (n.d.). "SPONGE: A generalized eigenproblem for clustering signed networks". In: *AISTATS 2019* ().
- Davenport, Mark A and Justin Romberg (2016). "An overview of low-rank matrix recovery from incomplete observations". In: *IEEE Journal of Selected Topics in Signal Processing* 10.4, pp. 608–622.
- Davis, Chandler and W. M. Kahan (1970). "The Rotation of Eigenvectors by a Perturbation. III". In: *SIAM Journal on Numerical Analysis* 7.1, pp. 1–46.
- Fei, Yingjie and Yudong Chen (2019). "Exponential error rates of SDP for block models: beyond Grothendieck's inequality". In: *IEEE Trans. Inform. Theory* 65.1, pp. 551–571. ISSN: 0018-9448.
- Fletcher, Roger (1981). "A nonlinear programming problem in statistics (educational testing)". In: *SIAM Journal on Scientific and Statistical Computing* 2.3, pp. 257–267.
- Gärtner, Bernd and Jiři Matoušek (2012). "Semidefinite programming". In: *Approximation Algorithms and Semidefinite Programming*. Springer, pp. 15–25.
- Giraud, Christophe and Nicolas Verzelen (2018). "Partial recovery bounds for clustering with the relaxed K-means". In: *Mathematical Statistics and Learning* 1.3, pp. 317–374.
- Goemans, Michel X (1997). "Semidefinite programming in combinatorial optimization". In: *Mathematical Programming* 79.1-3, pp. 143–161.

References III

- Goemans, Michel X and David P Williamson (1994). “New 34-approximation algorithms for the maximum satisfiability problem”. In: *SIAM Journal on Discrete Mathematics* 7.4, pp. 656–666.
- – (1995). “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. In: *Journal of the ACM (JACM)* 42.6, pp. 1115–1145.
- Grothendieck, Alexandre (1956). *Résumé de la théorie métrique des produits tensoriels topologiques*. Soc. de Matemática de São Paulo.
- Guédon, Olivier and Roman Vershynin (2016). “Community detection in sparse networks via Grothendieck’s inequality”. In: *Probability Theory and Related Fields* 165.3-4, pp. 1025–1049.
- Hajek, Bruce, Yihong Wu, and Jiaming Xu (2016). “Achieving exact cluster recovery threshold via semidefinite programming”. In: *IEEE Transactions on Information Theory* 62.5, pp. 2788–2797.
- He, Simai et al. (2008). “Semidefinite relaxation bounds for indefinite homogeneous quadratic optimization”. In: *SIAM Journal on Optimization* 19.2, pp. 503–523.
- Hegde, Chinmay, Aswin C Sankaranarayanan, and Richard G Baraniuk (2012). “Near-isometric linear embeddings of manifolds”. In: *2012 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, pp. 728–731.
- Hou, Jao Ping (2005). “Bounds for the least Laplacian eigenvalue of a signed graph”. In: *Acta Mathematica Sinica* 21.4, pp. 955–960.
- Karger, David, Rajeev Motwani, and Madhu Sudan (1998). “Approximate graph coloring by semidefinite programming”. In: *Journal of the ACM (JACM)* 45.2, pp. 246–265.

References IV

- Khot, Subhash and Assaf Naor (2009). “Approximate kernel clustering”. In: *Mathematika* 55.1-2, pp. 129–165.
- Koltchinskii, Vladimir (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Vol. 2033. Lecture Notes in Mathematics. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d’Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. Heidelberg: Springer, pp. x+254. ISBN: 978-3-642-22146-0. DOI: 10.1007/978-3-642-22147-7. URL: http://dx.doi.org/10.1007/978-3-642-22147-7.
- Kunegis, Jérôme et al. (2010). “Spectral analysis of signed graphs for clustering, prediction and visualization”. In: *SDM* 10, pp. 559–570.
- Lecué, Guillaume and Shahar Mendelson (2013). *Learning subgaussian classes: Upper and minimax bounds*. Tech. rep. CNRS, Ecole polytechnique and Technion.
- Lemaréchal, Claude, Arkadii Nemirovskii, and Yurii Nesterov (1995). “New variants of bundle methods”. In: *Mathematical programming* 69.1-3, pp. 111–147.
- Ma, Wing-Kin Ken (2010). “Semidefinite relaxation of quadratic optimization problems and applications”. In: *IEEE Signal Processing Magazine* 1053.5888/10.
- Mammen, Enno and Alexandre B. Tsybakov (1999). “Smooth discrimination analysis”. In: *Ann. Statist.* 27.6, pp. 1808–1829. ISSN: 0090-5364. DOI: 10.1214/aos/1017939240. URL: https://doi.org/10.1214/aos/1017939240.

References V

- Massart, Pascal (2007). *Concentration inequalities and model selection*. Vol. 1896. Lecture Notes in Mathematics. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. Berlin: Springer, pp. xiv+337. ISBN: 978-3-540-48497-4; 3-540-48497-3.
- Nesterov, Y (1997). "Semidefinite relaxation and non-convex quadratic optimization". In: *Optimization Methods and Software* 12, pp. 1–20.
- Olsson, Carl, Anders P Eriksson, and Fredrik Kahl (2007). "Solving large scale binary quadratic problems: Spectral methods vs. semidefinite programming". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.
- Peng, Jiming and Yu Wei (2007). "Approximating k-means-type clustering via semidefinite programming". In: *SIAM journal on optimization* 18.1, pp. 186–205.
- Peng, Jiming and Yu Xia (2005). "A new theoretical framework for k-means-type clustering". In: *Foundations and advances in data mining*. Springer, pp. 79–96.
- Pisier, Gilles (2012). "Grothendieck's theorem, past and present". In: *Bulletin of the American Mathematical Society* 49.2, pp. 237–323.
- Royer, Martin (2017). "Adaptive clustering through semidefinite programming". In: *Advances in Neural Information Processing Systems*, pp. 1795–1803.
- Scobey, P and DG Kabe (1978). "Vector quadratic programming problems and inequality constrained least squares estimation". In: *J. Indust. Math. Soc.* 28, pp. 37–49.

References VI

Shapiro, Alexander (1982). “Weighted minimum trace factor analysis”. In: *Psychometrika* 47.3, pp. 243–264.

Singer, Amit (2011). “Angular synchronization by eigenvectors and semidefinite programming”. In: *Applied and computational harmonic analysis* 30.1, pp. 20–36.

Sun, Jun et al. (2006). “The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem”. In: *SIAM review* 48.4, pp. 681–699.

Tsybakov, Alexandre B. (2003). “Optimal rate of aggregation”. In: *Computational Learning Theory and Kernel Machines (COLT-2003)*. Vol. 2777. Lecture Notes in Artificial Intelligence. Springer, Heidelberg, pp. 303–313.

Vershynin, Roman (2018). *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press.

Weinberger, Kilian Q, Benjamin Packer, and Lawrence K Saul (2005). “Nonlinear Dimensionality Reduction by Semidefinite Programming and Kernel Matrix Factorization.”. In: *AISTATS*. Vol. 2. 5. Citeseer, p. 6.

Weinberger, Kilian Q and Lawrence K Saul (2006a). “An introduction to nonlinear dimensionality reduction by maximum variance unfolding”. In: *AAAI*. Vol. 6, pp. 1683–1686.

– (2006b). “Unsupervised learning of image manifolds by semidefinite programming”. In: *International journal of computer vision* 70.1, pp. 77–90.

References VII

- Wolkowicz, Henry (1999). “Semidefinite and Lagrangian relaxations for hard combinatorial problems”. In: *IFIP Conference on System Modeling and Optimization*. Springer, pp. 269–309.
- Xu, M. et al. (2020). “Optimal Rates for Community Estimation in the Weighted Stochastic Block Model”. In: *Annals of statistics*.
- Yu, Y., T. Wang, and R. J. Samworth (2015). “A useful variant of the Davis–Kahan theorem for statisticians”. In: *Biometrika* 102.2, pp. 315–323. DOI: 10.1093/biomet/asv008. eprint: /oup/backfile/content_public/journal/biomet/102/2/10.1093_biomet_asv008/1/asv008.pdf. URL: http://dx.doi.org/10.1093/biomet/asv008.
- Zhang, Shuzhong (2000). “Quadratic maximization and semidefinite relaxation”. In: *Mathematical Programming* 87.3, pp. 453–465.