Optimization and learning techniques for clustering problems

Soledad Villar

Center for Data Science
Courant Institute of Mathematical Sciences

NEW YORK UNIVERSITY

Statistical Physics and Machine Learning back together
Cargèse, August 2018

Clustering

"Next AI revolution is unsupervised" Yann LeCun

Main task in unsupervised machine learning:

- Finding structure in unlabeled data.

k-means clustering

- Simple objective
- Useful for "generic data"

Clustering stochastic block model

- Specialized objective
- Algorithms get more precise but less robust to model changes

Data-driven clustering methods.

Clustering

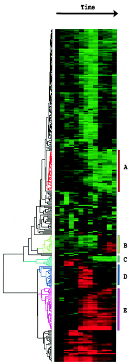
"Next AI revolution is unsupervised" Yann LeCun

Main task in unsupervised machine learning:

- Finding structure in unlabeled data.

Example:
Cluster analysis and display of genome-wide expression patterns.

- Gene expression as a function of time for different conditions.
- Clustering gene expression data groups together genes of known similar functionality.



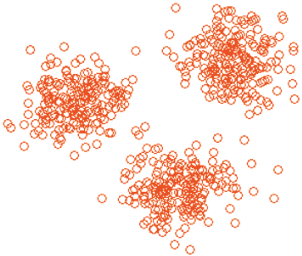
Eisen, Spellman, Brown, Botstein, PNAS, 1998

Summary

- k-means clustering
  - Landscape of manifold optimization
  - Lower bounds certificates from SDPs (data-driven)
- Clustering stochastic block model
  - AMP $\longrightarrow$ Spectral methods $\longrightarrow$ Graph neural networks
- Quadratic assignment

The k-means problem

Given a point cloud $\{x_i\}$, partition the points in clusters $C_1, \ldots, C_k$



k-means objective:

$$\min_{C_1,\ldots C_k} \sum_{t=1}^{k} \sum_{i \in C_t} \left\| x_i - \frac{1}{|C_t|} \sum_{j \in C_t} x_j \right\|^2$$

- NP-hard to minimize in general (even in the plane).

Lloyd's algorithm

1. Choose k centers at random.
2. Assign points to closest center.
3. Compute new centers are clusters centroids.

k-means++ is Lloyd's with smarter initialization.

Pros:

- Fast.
- Very easy to implement.
- Widely used and it works for most applications.

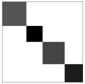
Cons:

- No guarantee of convergence (local minima).
- In extreme cases it may take exponentially many steps.
- Solutions depend heavily on initialization.
- Its output doesn't say how good of a solution it may be.

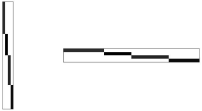
Optimization formulation

Taking $D_{ij} := \|x_i - x_j\|^2$, then

$$\sum_{t=1}^{k} \sum_{i \in C_t} \left\| x_i - \frac{1}{|C_t|} \sum_{j \in C_t} x_j \right\|^2 = \frac{1}{2} \operatorname{Tr}\left(D \underbrace{\sum_{t=1}^{k} \frac{1}{|C_t|} 1_{C_t} 1_{C_t}^\top} \right)$$



$=$



$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(D Y Y^\top) \\ \text{subject to} & Y \in \mathbb{R}^{n \times k} \\ & Y^\top Y = I_k \\ & Y Y^\top 1 = 1 \\ & Y \geq 0 \end{array}$$

Set of Y's is discrete.

Manifold optimization implementation

$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}(DYY^\top) + \lambda \|Y_-\|^2 \\ \text{subject to} & Y \in M = \{Y \in \mathbb{R}^{n\times k} : Y^\top Y \ = I_k, \quad YY^\top 1 = 1\} \end{array}$$

To implement the manifold optimization one needs:

$$\mathrm{grad}_f : M \to TM,$$
$$\mathrm{retr}_Y : T_Y M \to M,$$

where

$$\mathrm{retr}_Y(0) = 0, \quad \left.\frac{d}{dt}\right|_{t=0} \mathrm{retr}_Y(tV) = V.$$

Gradient descent:

$$Y_{n+1} = \mathrm{retr}_{Y_n}(-\alpha_n \, \mathrm{grad}_f(Y_n)).$$

Sepulchre, Absil, Mahony, 2008

Carson, Mixon, V., Ward, SampTA, 2017

Manifold optimization, retraction implementation

Homogeneous structure of M.
Let $O(1_n)$ be the $n \times n$ orthogonal matrices that fix 1.

$$\begin{aligned} M \times O(1_n) \times O(k) &\to M \\ (Y, Q, R) &\mapsto QYR \quad \text{(transitive action)} \end{aligned}$$

Multiplication of Y on the right by R is equivalent to multiplication on the left by $R' = YRY^\top \in O(1_n)$.

$V \in T_Y M$ can be written as $V = BY + YA$ where $A \in \mathfrak{so}(k)$, $B \in \mathfrak{so}(1_n)$ where BY and YA are orthogonal. We compute A, B explicitly.

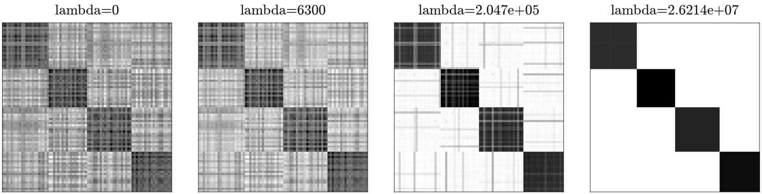
$$\begin{aligned} A &= Y^\top V \\ B &= VY^\top - YV^\top - 2YAY^\top \end{aligned}$$

$\mathrm{retr}_Y(V) = \exp(B)\exp(A')Y \in M$

Manifold optimization algorithm

$\lambda_0 \leftarrow 0$
repeat
  $Y_{n+1} \leftarrow \text{GradientDescent}(f_\lambda)$ {Initialized at Y_n}
  $\lambda_{n+1} \leftarrow 2\lambda_n + 1$
until $\|Y_-\|_F < \epsilon$

$$\begin{aligned} &\text{minimize}_Y \quad f_\lambda(Y) = \text{Tr}(DYY^\top) + \lambda \|Y_-\|^2 \\ &\text{subject to} \quad Y \in M = \{Y \in \mathbb{R}^{n\times k} : Y^\top Y = I_k, \quad YY^\top 1 = 1\} \end{aligned}$$



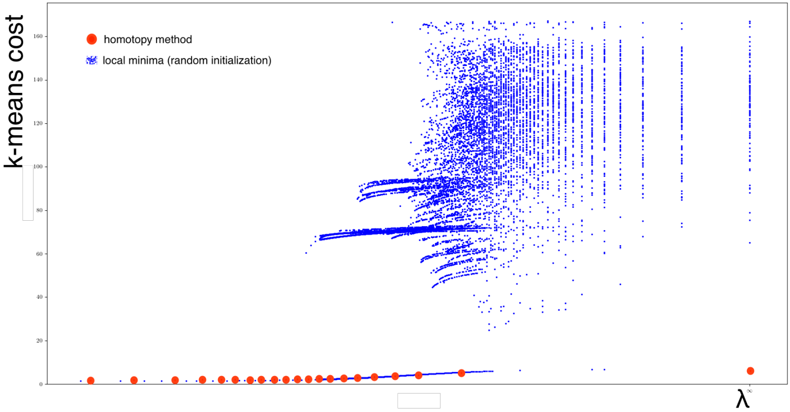
Problem structure allows to do each iteration in linear-time.

Carson, Mixon, V., Ward, SampTA, 2017

Claim: clustering is (maybe) not (that) hard

Optimization landscape

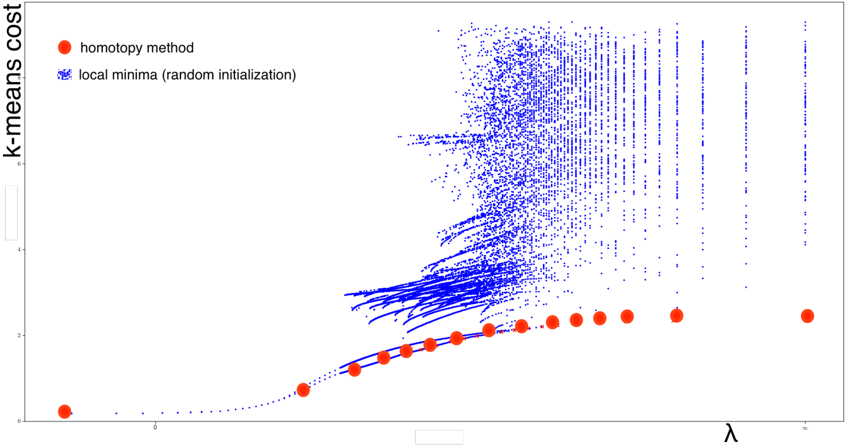
Points from three well-separated unit balls in small dimension.



Claim: clustering is (maybe) not (that) hard

No clustering structure

Points drawn from unit ball into three clusters in small dimension.

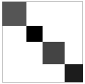


And it's optimal! How do I know that?

k-means optimization formulation

Recall $D_{ij} := \|x_i - x_j\|^2$, then

$$\sum_{t=1}^{k} \sum_{i \in C_t} \left\| x_i - \frac{1}{|C_t|} \sum_{j \in C_t} x_j \right\|^2 = \frac{1}{2} \operatorname{Tr}\left(D \underbrace{\sum_{t=1}^{k} \frac{1}{|C_t|} 1_{C_t} 1_{C_t}^\top} \right)$$



$=$

$$\begin{aligned} \text{minimize} \quad & \operatorname{Tr}(D Y Y^\top) \\ \text{subject to} \quad & Y \in \mathbb{R}^{n \times k} \\ & Y^\top Y = I_k \\ & Y Y^\top 1 = 1 \\ & Y \geq 0 \end{aligned}$$

Set of Y's is discrete.

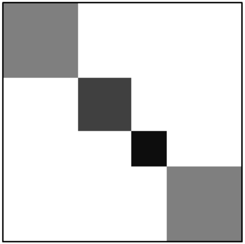
A semidefinite programming relaxation

Recall $D_{ij} := \|x_i - x_j\|^2$, then

$$\sum_{t=1}^{k} \sum_{i \in C_t} \left\| x_i - \frac{1}{|C_t|} \sum_{j \in C_t} x_j \right\|^2 = \frac{1}{2} \operatorname{Tr} \left(D \underbrace{\sum_{t=1}^{k} \frac{1}{|C_t|} 1_{C_t} 1_{C_t}^\top}_{} \right)$$

Relax to SDP:

$$\begin{array}{ll} \text{minimize} & \operatorname{Tr}(DX) \\ \text{subject to} & \operatorname{Tr}(X) = k \\ & X1 = 1 \\ & X \geq 0 \\ & X \succeq 0 \end{array}$$

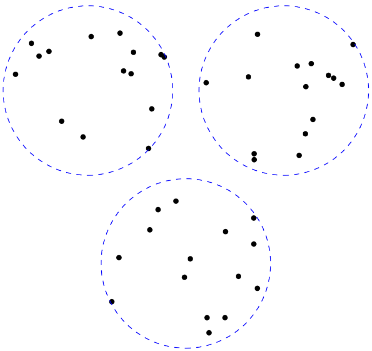


Peng, Wei, SIAM J. Optim., 2007

Stochastic ball model

$(\mathcal{D}, \gamma, n)$-**stochastic ball model**

- $\mathcal{D}$ = rotation-invariant distribution over unit ball in $\mathbb{R}^m$
- $\gamma_1, \ldots, \gamma_k$ = ball centers in $\mathbb{R}^m$
- Draw $r_{t,1}, \ldots, r_{t,n}$ i.i.d. from $\mathcal{D}$ for each $i \in \{1, \ldots, k\}$
- $x_{t,i} = \gamma_t + r_{t,i}$ = ith point from cluster t



Nellore, Ward, Information and Computation, 2015

Stochastic ball model

$(\mathcal{D}, \gamma, n)$-**stochastic ball model**

- $\mathcal{D}$ = rotation-invariant distribution over unit ball in $\mathbb{R}^m$
- $\gamma_1, \ldots, \gamma_k$ = ball centers in $\mathbb{R}^m$
- Draw $r_{t,1}, \ldots, r_{t,n}$ i.i.d. from $\mathcal{D}$ for each $i \in \{1, \ldots, k\}$
- $x_{t,i} = \gamma_t + r_{t,i}$ = ith point from cluster t

Nellore, Ward, Information and Computation, 2015

k-means SDP. *Stochastic ball model*

Exact recovery

SDP is tight for k unit balls in $\mathbb{R}^m$ with centers $\gamma_i \dots \gamma_k$

$$\min_{i \neq j} \|\gamma_i - \gamma_j\| \geq \min \left\{ 2\sqrt{2}(1 + \tfrac{1}{\sqrt{m}}),\ 2 + \tfrac{k^2}{m},\ 2 + \sqrt{\tfrac{2k}{m+2}} \right\}$$

Conjecture (Li, Li, Ling, Strohmer)

Phase transition for exact recovery (under stochastic ball model) at

$$\min_{i \neq j} \|\gamma_i - \gamma_j\| \geq 2 + \frac{c}{m}$$

Awasthi, Bandeira, Charikar, Krishnaswamy, V., Ward, Proc. ITCS, 2015

Iguchi, Mixon, Peterson, V., Mathematical Programming, 2016

Li, Li, Ling, Strohmer, `arXiv:1710.06008` 2017

k-means SDP. *Subgaussian mixtures*

Relax and "round" (weak recovery).

Theorem

Centers: $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$. Estimated centers $v_1, \ldots, v_k$

$$\frac{1}{k}\sum_{i=1}^{k} \|v_i - \hat{\gamma}_i\|^2 \lesssim k^2\sigma^2 \text{ whp provided } \min_{i \neq j} \|\gamma_i - \gamma_j\| \gtrsim k\sigma.$$

Example: `http://solevillar.github.io/2016/07/05/Clustering-MNIST-SDP.html`

Mixon, V., Ward, Information and Inference, 2017

Proof technique: Guédon, Vershynin, Probability Theory and Related Fields, 2016

SDPs are slow!

Fast lower bound certificates

Want: Lower bound on k-means value,

Given $\{x_i\}_{i\in T} \subseteq \mathbb{R}^m$, let $W=$ value of k-means++ initialization

$$\mathrm{val}_{\mathrm{kmeans}}(\mathrm{T}) := \min \frac{1}{|T|} \sum_{t=1}^{k} \sum_{i\in C_t} \left\| x_i - \frac{1}{|C_t|} \sum_{j\in C_t} x_j \right\|^2 \geq \frac{1}{8(\log k + 2)} \mathbb{E} W$$

Running k-means++ on MNIST training set with $k = 10$ gives:

- Upper bound (by running the algorithm): 39.22
- Lower bound (estimated from above): 2.15

Can we get a better lower bound?

Arthur, Vassilvitskii, 2007

Idea

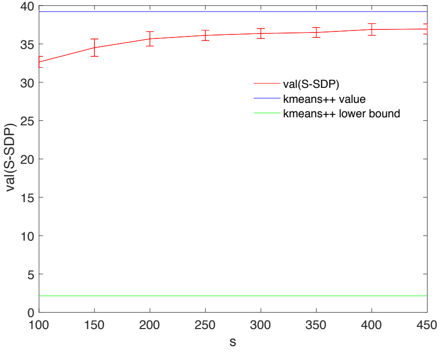
Given $\{x_i\}_{i \in T} \subseteq \mathbb{R}^m$, draw $S \sim \mathrm{Unif} \in \binom{T}{s}$.

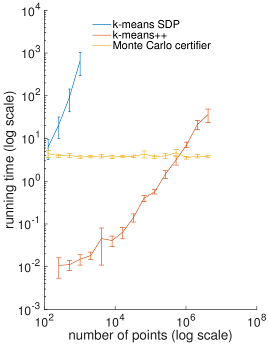
$$\begin{aligned}
\mathbb{E}\,\mathrm{val}_{\mathrm{SDP}}(\mathrm{S}) &\leq \mathbb{E}\,\mathrm{val}_{\mathrm{kmeans}}(\mathrm{S}) \\
&\leq \mathbb{E}\left[\frac{1}{s}\sum_{t=1}^{k}\sum_{i \in C_t^* \cap S}\left\|x_i - \frac{1}{|C_t^* \cap S|}\sum_{j \in C_t^* \cap S} x_j\right\|^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{s}\sum_{t=1}^{k}\sum_{i \in C_t^* \cap S}\left\|x_i - \frac{1}{|C_t^*|}\sum_{j \in C_t^*} x_j\right\|^2\right] = \mathrm{val}_{\mathrm{kmeans}}(\mathrm{T})
\end{aligned}$$

Goal: Establish $\mathbb{E}\,\mathrm{val}_{\mathrm{SDP}}(\mathrm{S}) > B$ with small p-value.

Mixon, V., arXiv:1710.00956

Better performance, and fast!





- Taking $s = 100 \ll 60,000$ MNIST points already works
- Constant runtime in N, faster than k-means++ for $N \gtrsim 10^6$.
- Theorem for mixtures of Gaussians (additive bound)

Mixon, V., arXiv:1710.00956

Generic algorithms vs. model-dependent algorithms

Clustering the stochastic block model

- Sparse graph clustering (non-geometric)
- Similar generic algorithmic (SDP, manifold optimization)
- Model-tailored algorithms (AMP, spectral methods on Non-bactracking matrices and Bethe Hessian)
- Can we learn the algorithm from data?
  - Use graph neural networks

Decelle, Krzakala, Moore, Zdeborová, 2013

Krzakala, Moore, Mossel, Neeman, Sly, Zdeborová, Zhang, 2013

Saade, Krzakala, Zdeborová, 2014

Abbe, 2018

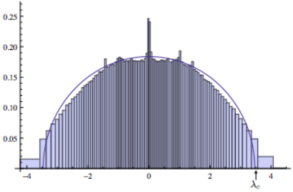
Spectral redemption

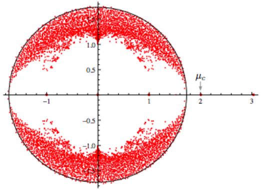
$A \sim SBM(a/n, b/n, n, 2)$ sparse.

Spectrum doesn't concentrate (high degree vertices dominate it)
Laplacian is not useful for clustering

Consider the non-backtracking operator (from linearized BP)

$$B_{(i \to j)(i' \to j')} = \begin{cases} 1 \text{ if } j = i' \text{ and } j' \neq i \\ 0 \text{ otherwise} \end{cases}$$





Second eigenvector of B reveals clustering structure

Krzakala, Moore, Mossel, Neeman, Sly, Zdeborová, Zhang, 2013

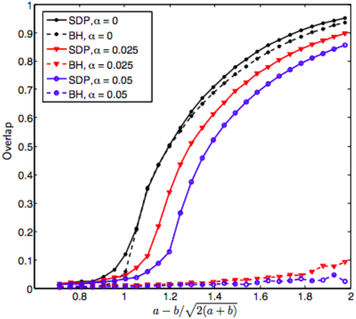
Bethe Hessian

$$BH(r) = (r^2 - 1)I - rA + D$$

Fixed points of BP $\longleftrightarrow$ Stationary points of Bethe free energy

Second eigenvector reveals clustering structure

Pitfall: highly dependent in the model



Goal: Combine graph operators $I, D, A, \ldots$ to generate robust "data-driven spectral methods" for problems in graphs

Saade, Krzakala, Zdeborová, 2014

Javanmard, Montanari, Ricci-Tersenghi, 2015

Graph neural networks

Power method: $v^{t+1} = Mv^t \qquad t = 1, \dots, T.$

Graph with adjacency matrix A. Set $\mathcal{M} = \{I_n, D, A, A^2, \dots A^{2^J}\}$,

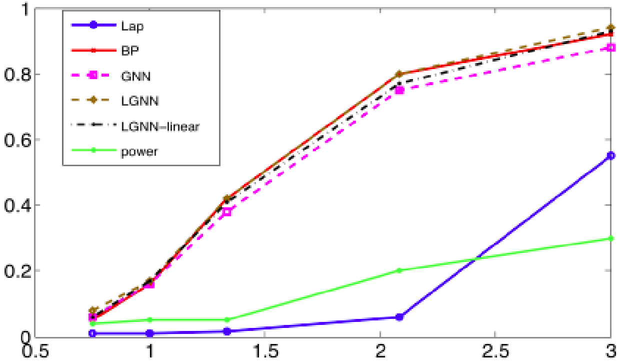
$$v_l^{t+1} = \rho\left(\sum_{M \in \mathcal{M}} M v^t {\theta_{M,l}}^t\right) \,, \, l = 1 \dots d_{t+1}$$

with $v^t \in \mathbb{R}^{n \times d_t}$,
$\Theta = \{\theta_1^t, \dots, \theta_{|\mathcal{M}|}^t\}_t$, $\theta_M^t \in \mathbb{R}^{d_t \times d_{t+1}}$ trainable parameters.

- Extension to line graph (GNN with non-backtracking).
- Covariant wrt permutations $G \mapsto \phi(G)$ then $G_\Pi \mapsto \Pi\phi(G)$.
- Preliminary theoretical analysis of energy landscape.
  - Strong assumptions $\Rightarrow$ local minima have low energy.

Scarselli, Tsoi, Hagenbuchner, Monfardini, IEEE Transactions on Neural Networks, 2009
Chen, Li, Bruna, 2017

Numerical performance. SBM $k = 2$

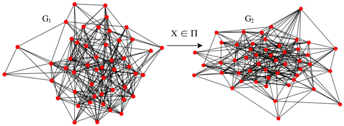


Overlap as function of SNR

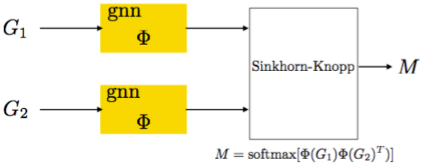
Chen, Li, Bruna, 2017

Extension to quadratic assignment

Graph matching: $\min_{X \in \Pi} \|G_1 X - X G_2\|_F^2 = \|G_1 X\|^2 + \|X G_2\|^2 - 2\langle G_1 X, X G_2 \rangle$



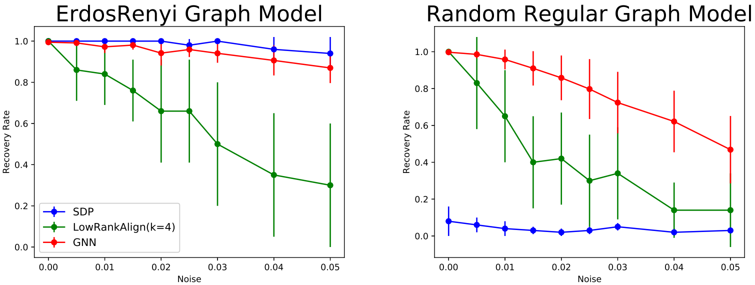
Siamese neural network:



$G_2 = \pi G_1 + N$ $\quad$ $N \sim$Erdos-Renyi
$G_1 \sim$ Erdos-Renyi
$G_1 \sim$ Random regular

Nowak, V., Bandeira, Bruna, 2017

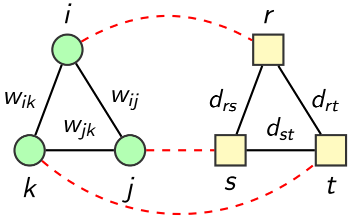
Numerical experiment



Our model runs in $O(n^2)$, LowRankAlign is $O(n^3)$, SDP in $O(n^4)$.

Nowak, V., Bandeira, Bruna, 2017

Quadratic Assignment Problem (QAP)



$$\min_{\pi} \quad \sum_{i\in A}\sum_{j\in A}\sum_{r\in B}\sum_{s\in B} \underbrace{w_{ij}\, d_{rs}}_{E_{ijrs}} \pi_{ir}\pi_{js}$$

$$\text{subject to} \quad \sum_{r} \pi_{ir} = 1 \quad \forall i \in A$$

$$\sum_{i} \pi_{ir} = 1 \quad \forall r \in B$$

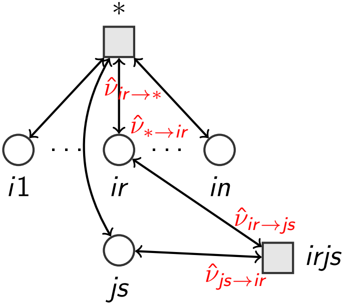
$$\pi_{ir} \in \{0, 1\}$$

Onaran, Villar, ongoing work

Max-Product Belief Propagation for QAP

- Posterior probability:

$$p(\pi) = \frac{1}{Z} \prod_i \mathbb{1}\{\sum_t \pi_{it} = 1\} \prod_r \mathbb{1}\{\sum_k \pi_{kr} = 1\} \prod_{j \neq i} \prod_{s \neq r} e^{-\beta \pi_{ir} \pi_{js} E_{ijrs}}$$



Max-Product Message Updates

- Variable to check nodes:

$$\nu_{ir\to *}(1) \cong \prod_{kt} e^{-\beta \hat{\nu}_{kt\to ir}(1) E_{irkt}}$$

$$\nu_{ir\to *}(0) \cong 1$$

- Check to variable nodes:

$$\nu_{*\to ir}(1) \cong \nu_{ir\to *}(1) \prod_{\ell \neq i} \nu_{\ell r\to *}(0) \prod_{t\neq r} \nu_{it\to *}(0)$$

$$\nu_{*\to ir}(0) \cong \nu_{ir\to *}(0) \max_{t\neq r, \ell\neq i} \nu_{it\to *}(1)\nu_{\ell r\to *}(1) \prod_{u\neq t,r} \nu_{iu\to *}(0) \prod_{m\neq \ell, i} \nu_{mr\to *}(0)$$

- Variable back to variable nodes:

$$\hat{\nu}_{ir\to js}(1) \cong \nu_{*\to ir}(1) \prod_{ku\neq js} e^{-\beta \hat{\nu}_{ku\to ir}(1) E_{irku}}$$

$$\hat{\nu}_{ir\to js}(0) \cong \nu_{*\to ir}(0)$$

Simplifications (Min-Sum Algorithm)

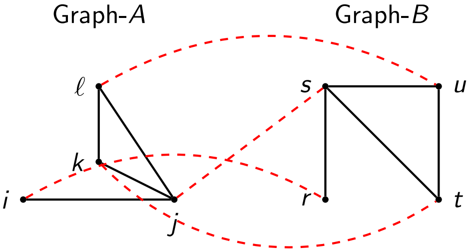
Introduce log-likelihood ratio:

$$\hat{\ell}_{ir\to js} = \frac{1}{\beta} \log \frac{\hat{\nu}_{ir\to js}(1)}{\hat{\nu}_{ir\to js}(0)}$$

Then the algorithm simplifies to one update step as follows,

$$\begin{aligned}\hat{\ell}_{ir\to js} \leftarrow \min_{t\neq r, l\neq i} & \sum_{ku} \frac{e^{\beta\hat{\ell}_{ku\to it}}}{1+e^{\beta\hat{\ell}_{ku\to it}}} E_{itku} + \frac{e^{\beta\hat{\ell}_{ku\to \ell r}}}{1+e^{\beta\hat{\ell}_{ku\to \ell r}}} E_{\ell rku} \\ & -2\sum_{ku\neq js} \frac{e^{\beta\hat{\ell}_{ku\to ir}}}{1+e^{\beta\hat{\ell}_{ku\to ir}}} E_{irku} + \frac{e^{\beta\hat{\ell}_{js\to ir}}}{1+e^{\beta\hat{\ell}_{js\to ir}}} E_{irjs}\end{aligned}$$

A special case: Graph Matching (Isomorphism)



When $A \sim \mathrm{ER}(n, 0.5)$ and

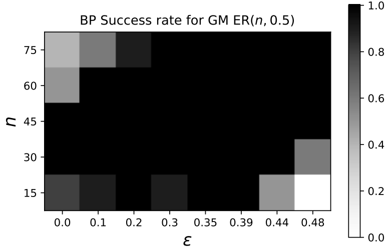
$$B = PAP^T \oplus Z$$

for some $P \in \Pi$ where $Z \sim \mathrm{Ber}(\epsilon)$. Find P given unlabeled graphs.

$$\min_{P} \quad \|A - PBP^T\|_F$$

$$\text{subject to} \quad P \in \Pi$$

equivalent to QAP with $E_{ijrs} = A_{ij} \oplus B_{rs}$.



Summary

- k-means clustering
  - Nice manifold optimization landscape
  - Lower bounds certificates from SDPs (data-driven)
- Clustering stochastic block model
  - AMP $\longrightarrow$ Spectral methods $\longrightarrow$ Graph neural networks
- Quadratic assignment

Thank you

Relax no need to round: Integrality of clustering formulations
P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, R. Ward
arXiv:1408.4045, Innovations in Theoretical Computer Science (ITCS) 2015

Probably certifiably correct *k*-means clustering
T. Iguchi, D. G. Mixon, J. Peterson, S. Villar
arXiv:1509.07983, Mathematical Programming, 2016

Clustering subgaussian mixtures by semidefinite programming
D. G. Mixon, S. Villar, R. Ward
arXiv:1602.06612, Information and Inference: A Journal of the IMA, 2017

Monte Carlo approximation certificates for k-means clustering
D. G. Mixon, S. Villar
arXiv:1710.00956

Supervised Community Detection with Hierarchical Graph Neural Networks
Z. Chen, L. Li, J. Bruna
arXiv:1705.08415

A note on learning algorithms for quadratic assignment with Graph Neural Networks
A. Nowak, S. Villar, A. S. Bandeira, J. Bruna, ICML (PADL) 2017
arXiv:1706.07450