Optimal Non-Convex Exact Recovery in Stochastic Block Model via Projected Power Method

Peng Wang

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong (CUHK)

(Joint Work with Huikang Liu, Zirui Zhou, and Anthony Man-Cho So)

Proceedings of the 38th International Conference on Machine Learning
(ICML 2021)

June 20, 2021

Outline

Symmetric Stochastic Block Model (SBM)

- Model setup
  - (Ground Truth) Let $\boldsymbol{H}^* \in \{0,1\}^{n \times K}$ denote a clustering matrix representing a partition of a vertex set V of n nodes into K equal-sized communities.

Symmetric Stochastic Block Model (SBM)

- Model setup
  - (Ground Truth) Let $\boldsymbol{H}^* \in \{0,1\}^{n\times K}$ denote a clustering matrix representing a partition of a vertex set V of n nodes into K equal-sized communities.
  - (Observed Graph) A graph G has the vertex set V and the elements $\{a_{ij}\}_{1\le i\le j\le n}$ of its adjacency matrix $\boldsymbol{A}$ is generated independently as follows:
    - If vertices i, j belong to the same community, i.e., ${\boldsymbol{h}_i^*}^T \boldsymbol{h}_j^* = 1$, they are connected with probability p, i.e.,

$$a_{ij} = \begin{cases} 1, & \text{w.p. } p, \\ 0, & \text{w.p. } 1-p, \end{cases} \quad (a_{ij} \sim \textbf{Bern}(p)).$$

    - If i, j belong to different communities, i.e., ${\boldsymbol{h}_i^*}^T \boldsymbol{h}_j^* = 0$,

$$a_{ij} \sim \textbf{Bern}(q),$$

  where $\boldsymbol{h}_i^*$ denotes the i-th row of $\boldsymbol{H}^*$, $p, q \in [0,1]$, and $p > q$.

Symmetric Stochastic Block Model (SBM)

- Model setup
  - (Ground Truth) Let $\boldsymbol{H}^* \in \{0,1\}^{n\times K}$ denote a clustering matrix representing a partition of a vertex set V of n nodes into K equal-sized communities.
  - (Observed Graph) A graph G has the vertex set V and the elements $\{a_{ij}\}_{1\leq i\leq j\leq n}$ of its adjacency matrix $\boldsymbol{A}$ is generated independently as follows:
    - If vertices i, j belong to the same community, i.e., ${\boldsymbol{h}_i^*}^T \boldsymbol{h}_j^* = 1$, they are connected with probability p, i.e.,

$$a_{ij} = \begin{cases} 1, & \text{w.p. } p, \\ 0, & \text{w.p. } 1-p, \end{cases} \quad (a_{ij} \sim \mathbf{Bern}(p)).$$

    - If i, j belong to different communities, i.e., ${\boldsymbol{h}_i^*}^T \boldsymbol{h}_j^* = 0$,

$$a_{ij} \sim \mathbf{Bern}(q),$$

  where $\boldsymbol{h}_i^*$ denotes the i-th row of $\boldsymbol{H}^*$, $p, q \in [0,1]$, and $p > q$.
- (Exact Recovery) Recover the underlying communities exactly, i.e., $\boldsymbol{H}^*\boldsymbol{Q}$ for any $\boldsymbol{Q} \in \Pi_K$, with high probability.

Maximum Likelihood (ML) Formulation

- According to **[Amini et al., 2018]**, the ML estimator of $\boldsymbol{H}^*$ in the symmetric SBM is the solution of

$$\max\left\{\langle \boldsymbol{H}, \boldsymbol{A}\boldsymbol{H}\rangle : \boldsymbol{H}\mathbf{1}_K = \mathbf{1}_n, \boldsymbol{H}^T\mathbf{1}_n = m\mathbf{1}_K, \boldsymbol{H} \in \{0,1\}^{n\times K}\right\}.$$

  - $\boldsymbol{H}\mathbf{1}_K = \mathbf{1}_n$ requires each vertex to belong to only one cluster.
  - $\boldsymbol{H}^T\mathbf{1}_n = m\mathbf{1}_K$ requires all clusters to be of equal size, where $m = n/K$ is the cluster size.
  - The objective is to maximize the number of within-cluster edges.
  - NP-hard in the worst-case.

Maximum Likelihood (ML) Formulation

- According to **[Amini et al., 2018]**, the ML estimator of $\boldsymbol{H}^*$ in the symmetric SBM is the solution of

$$\max\left\{\langle \boldsymbol{H}, \boldsymbol{A}\boldsymbol{H}\rangle : \boldsymbol{H}\mathbf{1}_K = \mathbf{1}_n, \boldsymbol{H}^T\mathbf{1}_n = m\mathbf{1}_K, \boldsymbol{H} \in \{0,1\}^{n\times K}\right\}.$$

  - $\boldsymbol{H}\mathbf{1}_K = \mathbf{1}_n$ requires each vertex to belong to only one cluster.
  - $\boldsymbol{H}^T\mathbf{1}_n = m\mathbf{1}_K$ requires all clusters to be of equal size, where $m = n/K$ is the cluster size.
  - The objective is to maximize the number of within-cluster edges.
  - NP-hard in the worst-case.
- Logarithmic sparsity regime of the SBM, i.e.,

$$p = \frac{\alpha \log n}{n},\ q = \frac{\beta \log n}{n}$$

  for some constants $\alpha > \beta > 0$.
- **Fact [Abbe and Sandon, 2015].** In the symmetric SBM, exact recovery is impossible if $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{K}$, while it is possible if $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{K}$ (the information-theoretic limit).

Outline

Projected Power Method (PPM) with Mild Initialization

- Let $\mathcal{H} = \{\boldsymbol{H} \in \mathbb{R}^{n\times K} : \boldsymbol{H}\mathbf{1}_K = \mathbf{1}_n, \boldsymbol{H}^T\mathbf{1}_n = m\mathbf{1}_K, \boldsymbol{H} \in \{0,1\}^{n\times K}\}$. For any $\boldsymbol{C} \in \mathbb{R}^n$, let

$$\mathcal{T}(\boldsymbol{C}) = \text{arg min}\left\{\|\boldsymbol{H} - \boldsymbol{C}\|_F : \boldsymbol{H} \in \mathcal{H}\right\}. \tag{1}$$

- **Proposition.** Problem (1) is equivalent to a minimum-cost assignment problem, which can be solved in $\mathcal{O}(K^2 n \log n)$ time.

- The projected power iterations take the form

$$\boldsymbol{H}^{k+1} \in \mathcal{T}(\boldsymbol{A}\boldsymbol{H}^k), \text{ for all } k \geq 1. \tag{2}$$

- Initialization condition

$$\boldsymbol{H}^0 \in \mathbb{M}_{n,K} \text{ s.t. } \min_{\boldsymbol{Q}\in\Pi_K} \|\boldsymbol{H}^0 - \boldsymbol{H}^*\boldsymbol{Q}\|_F \lesssim \theta\sqrt{n}, \tag{3}$$

where θ is a specified constant and $\mathbb{M}_{n,K}$, Π_K denotes the collection of all clustering matrices and all $K \times K$ permutation matrices, respectively.

Master Theorem (Informal)

- **Theorem.** Suppose that the following hold:
  (i) (Data input) Let $\boldsymbol{A} \sim \mathrm{SBM}(\boldsymbol{H}^*, n, K, p, q)$.
  (ii) (Degree requirement) $p = \alpha \log n/n$, $q = \beta \log n/n$, and $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{K}$.
  (iii) (Sampling requirement) n is sufficiently large.

  The following statement holds with probability at least $1 - n^{-\Omega(1)}$: If the initial point $\boldsymbol{H}^0$ satisfies the partial recovery condition in (3) with a proper θ, PPM outputs a true partition in $\mathcal{O}(\log n/\log\log n)$ projected power iterations.

Master Theorem (Informal)

- **Theorem.** Suppose that the following hold:
  - (i) (Data input) Let $\boldsymbol{A} \sim \mathrm{SBM}(\boldsymbol{H}^*, n, K, p, q)$.
  - (ii) (Degree requirement) $p = \alpha \log n/n$, $q = \beta \log n/n$, and $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{K}$.
  - (iii) (Sampling requirement) n is sufficiently large.

  The following statement holds with probability at least $1 - n^{-\Omega(1)}$: If the initial point $\boldsymbol{H}^0$ satisfies the partial recovery condition in (3) with a proper θ, PPM outputs a true partition in $\mathcal{O}(\log n/\log\log n)$ projected power iterations.

- **Corollary.** Consider the same setting as above. It holds with probability at least $1 - n^{-\Omega(1)}$ that PPM outputs a true partition in $\mathcal{O}(n\log^2 n/\log\log n)$ time.

Comments on the Master Theorem

- While the ML formulation is NP-hard in the worst case, the assumption that $\boldsymbol{A}$ arises from the symmetric SBM allows us to conduct an average-case analysis.
- The total time complexity of the proposed method is nearly-linear, which is competitive with those of the most efficient methods in the literature.

Thank You!

References I

Abbe, E. and Sandon, C. (2015).
Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery.
In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE.

Amini, A. A., Levina, E., et al. (2018).
On semidefinite relaxations for the block model.
The Annals of Statistics, 46(1):149–179.

Boumal, N. (2016).
Nonconvex Phase Synchronization.
SIAM Journal on Optimization, 26(4):2355–2377.

Candès, E. J., Li, X., and Soltanolkotabi, M. (2015).
Phase retrieval via Wirtinger flow: Theory and algorithms.
IEEE Transactions on Information Theory, 61(4):1985–2007.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2019).
Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval.
Mathematical Programming, 176(1-2):5–37.

Chi, Y., Lu, Y. M., and Chen, Y. (2019).
Nonconvex optimization meets low-rank matrix factorization: An overview.
IEEE Transactions on Signal Processing, 67(20):5239–5269.

Fei, Y. and Chen, Y. (2018).
Exponential error rates of sdp for block models: Beyond grothendieck's inequality.
IEEE Transactions on Information Theory, 65(1):551–571.

References II

Gao, C., Ma, Z., Zhang, A. Y., and Zhou, H. H. (2017).
Achieving optimal misclassification proportion in stochastic block models.
The Journal of Machine Learning Research, 18(1):1980–2024.

Li, X., Zhu, Z., Man-Cho So, A., and Vidal, R. (2020).
Nonconvex robust low-rank matrix recovery.
SIAM Journal on Optimization, 30(1):660–686.

Liu, H., Yue, M.-C., and So, A. M.-C. (2020).
A unified approach to synchronization problems over subgroups of the orthogonal group.
arXiv preprint arXiv:2009.07514.

McSherry, F. (2001).
Spectral partitioning of random graphs.
In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE.

Su, L., Wang, W., and Zhang, Y. (2019).
Strong consistency of spectral clustering for stochastic block models.
IEEE Transactions on Information Theory, 66(1):324–338.

Yun, S.-Y. and Proutiere, A. (2016).
Optimal cluster recovery in the labeled stochastic block model.
In *Advances in Neural Information Processing Systems*, pages 965–973.

Zhang, Y., Qu, Q., and Wright, J. (2020).
From symmetry to geometry: Tractable nonconvex problems.
arXiv preprint arXiv:2007.06753.