Fundamental limits of symmetric low-rank matrix estimation

GDR MEGA, Paris

November, 2017

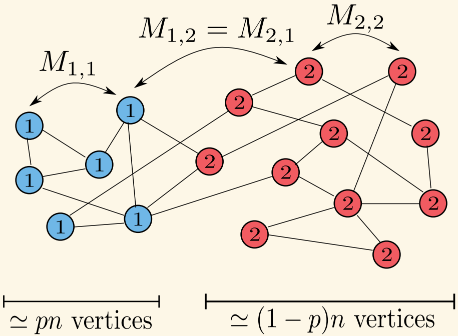
Francesco Caltagirone, Marc Lelarge & Léo Miolane

Community detection

The Stochastic Block Model (SBM)

$\mathbf{G}$ is generated as follows:

- n vertices: $1, \ldots, n$.
- Each vertex i has a label $X_i \in \{1, 2\}$ where $(X_k)_k \overset{\text{i.i.d.}}{\sim} 1 + \mathrm{Ber}(1-p)$.
- Two vertices i, j are then connected with probability M_{X_i, X_j}.

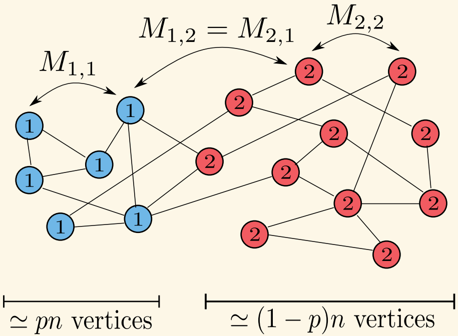


Community detection

The Stochastic Block Model (SBM)

$\mathbf{G}$ is generated as follows:

- n vertices: $1, \ldots, n$.
- Each vertex i has a label $X_i \in \{1, 2\}$ where $(X_k)_k \overset{\text{i.i.d.}}{\sim} 1 + \mathrm{Ber}(1-p)$.
- Two vertices i, j are then connected with probability M_{X_i, X_j}.



- **Goal:** given the graph $\mathbf{G}$ we want to recover the labels $\mathbf{X}$.
- **Weak Reconstruction:** Estimate $\mathbf{X}$ better than a "random guess".

Setting

- The connectivity matrix will be of the form:

$$\mathbf{M} = \frac{d}{n} \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

$a, c > b$ and $pa + (1-p)b = pb + (1-p)c = 1$.

Setting

- The connectivity matrix will be of the form:

$$\mathbf{M} = \frac{d}{n}\begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

  $a, c > b$ and $pa + (1-p)b = pb + (1-p)c = 1$.
- **Important quantity:** the signal-to-noise ratio

$$\lambda = d(1-b)^2$$

Setting

- The connectivity matrix will be of the form:

$$\mathbf{M} = \frac{d}{n}\begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

  $a, c > b$ and $pa + (1-p)b = pb + (1-p)c = 1$.
- **Important quantity:** the signal-to-noise ratio

$$\lambda = d(1-b)^2$$

Mossel et al., 2015, Massoulié, 2014, Mossel et al., 2013

In the case of two symmetric communities ($p = 1/2$), when $d > 1$ is fixed and $n \to \infty$,

- if $\lambda \leq 1$ it is not possible to recover the partition $\mathbf{X}$ better that a "random guess".
- if $\lambda > 1$ it is possible to recover the labels better than chance.

Asymmetric communities

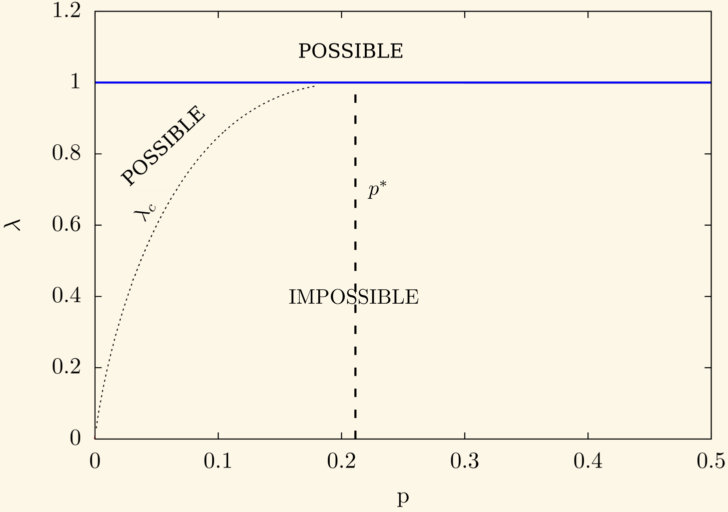
The main picture

- Does this phase transition at $\lambda = 1$ still hold when $p < 1/2$?
- The physicist's conjecture for the large degree limit ($d \to \infty$):

Asymmetric communities

The main picture

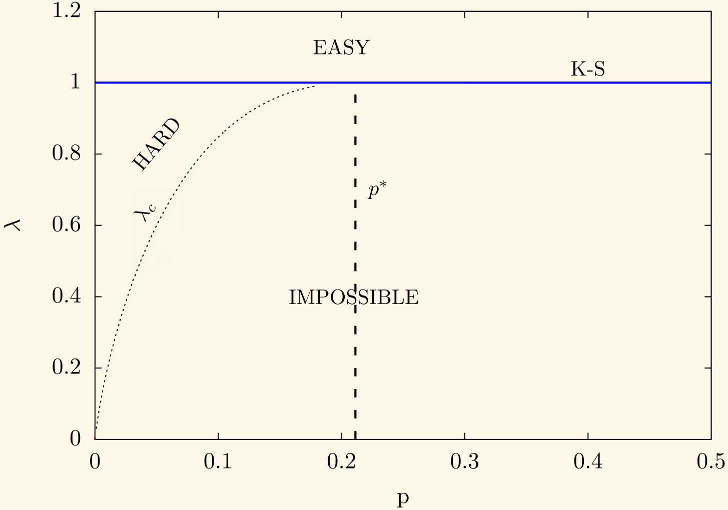
- Does this phase transition at $\lambda = 1$ still hold when $p < 1/2$?
- The physicist's conjecture for the large degree limit ($d \to \infty$):



Asymmetric communities

The main picture

- Does this phase transition at $\lambda = 1$ still hold when $p < 1/2$?
- The physicist's conjecture for the large degree limit ($d \to \infty$):

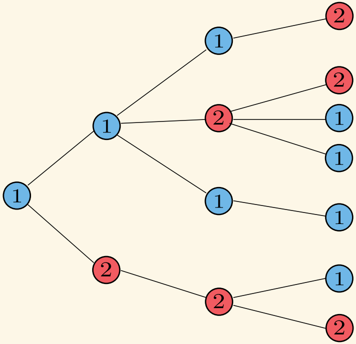


Part 1.

Local weak convergence of the SBM

Local weak convergence of the SBM

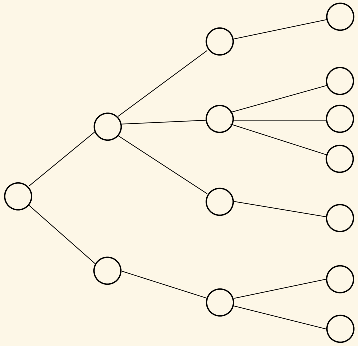
The Stochastic Block Model converges locally weakly to a "Labeled Poison Galton-Watson tree".



- Offspring distribution: Pois(d).
- The labels "propagate" from the root according to the transition matrix $\begin{pmatrix} pa & (1-p)b \\ pb & (1-p)c \end{pmatrix}$

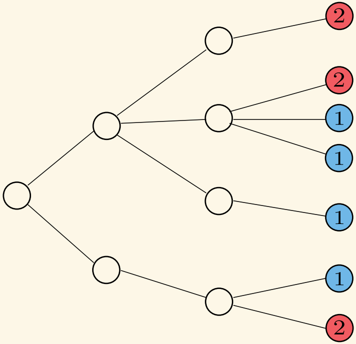
Reconstruction on trees

- **An issue:** the Galton-Watson tree, without the labels, does not give any information about the label of the root!



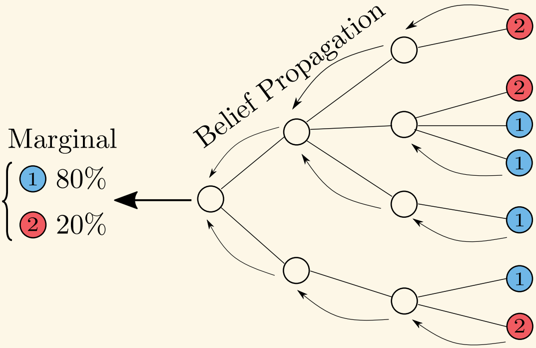
Reconstruction on trees

- **An issue:** the Galton-Watson tree, without the labels, does not give any information about the label of the root!
- We thus suppose that the labels at depth r are revealed. Can we infer the label of the root as $r \to \infty$?



Reconstruction on trees

- **An issue:** the Galton-Watson tree, without the labels, does not give any information about the label of the root!
- We thus suppose that the labels at depth r are revealed. Can we infer the label of the root as $r \to \infty$?



- Belief-Propagation gives the marginal distribution of the root given $\mathbf{G}$ and the labels at depth r.

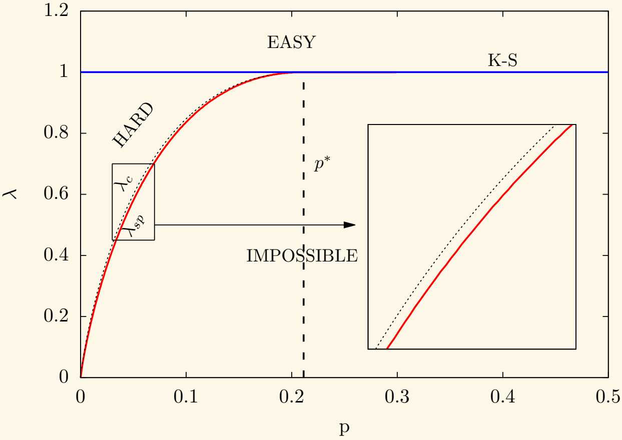
An impossibility result

- Studying the “BP recursion” one see that when $\lambda < \lambda_{\mathrm{sp}}$, the marginal does not contain any information about the true label.

An impossibility result

- Studying the "BP recursion" one see that when $\lambda < \lambda_{\mathrm{sp}}$, the marginal does not contain any information about the true label.

We thus obtain the "impossibility curve" $\lambda_{sp}(p)$ below:



Part 2.

Low-rank matrix estimation

Low-rank matrix estimation

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \mathsf{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $$\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{if } X_k = 1 \\ -\sqrt{p/(1-p)} & \text{if } X_k = 2 \end{cases}.$$

[1] Yash Deshpande and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.

Low-rank matrix estimation

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \mathsf{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $$\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{if } X_k = 1 \\ -\sqrt{p/(1-p)} & \text{if } X_k = 2 \end{cases}.$$

The Bernoulli noise model (1) is "equivalent" to the Gaussian noise model (when $n, d \to \infty$)[1]:

$$A'_{i,j} = \frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j + \sqrt{\frac{d}{n}}Z_{i,j} \tag{2}$$

where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$,

[1] Yash Deshpande and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.

Low-rank matrix estimation

From Bernoulli to Gaussian noise

$$A_{i,j} \sim \text{Ber}\left(\frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j\right) \tag{1}$$

where $$\tilde{X}_k = \begin{cases} \sqrt{(1-p)/p} & \text{if } X_k = 1 \\ -\sqrt{p/(1-p)} & \text{if } X_k = 2 \end{cases}.$$

The Bernoulli noise model (1) is "equivalent" to the Gaussian noise model (when $n, d \to \infty$)[1]:

$$A'_{i,j} = \frac{d}{n} + \frac{\sqrt{d}\sqrt{\lambda}}{n}\tilde{X}_i\tilde{X}_j + \sqrt{\frac{d}{n}}Z_{i,j} \tag{2}$$

where $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and thus to

$$Y_{i,j} = \sqrt{\frac{\lambda}{n}}\tilde{X}_i\tilde{X}_j + Z_{i,j}$$

[1] Yash Deshpande and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.

Low-rank matrix estimation

The new statistical model

"Spiked Wigner" model

$$\underbrace{\mathbf{Y}}_{\text{observations}} = \sqrt{\frac{\lambda}{n}} \underbrace{\mathbf{X}\mathbf{X}^\intercal}_{\text{signal}} + \underbrace{\mathbf{Z}}_{\text{noise}}$$

- $\mathbf{X}$: vector of dimension n with entries $X_i \overset{\text{i.i.d.}}{\sim} P_0$. $\mathbb{E}X_1 = 0$, $\mathbb{E}X_1^2 = 1$.
- $Z_{i,j} = Z_{j,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.
- λ: signal-to-noise ratio.
- λ and P_0 are known by the statistician.

Goal: recover the low-rank matrix $\mathbf{X}\mathbf{X}^\intercal$ from $\mathbf{Y}$.

Principal component analysis (PCA)

B.B.P. phase transition

Spectral estimator:
Estimate $\mathbf{X}$ using the eigenvector $\hat{\mathbf{x}}_n$ associated with the largest eigenvalue μ_n of $\mathbf{Y}/\sqrt{n}$.

Principal component analysis (PCA)

B.B.P. phase transition

Spectral estimator:
Estimate $\mathbf{X}$ using the eigenvector $\hat{\mathbf{x}}_n$ associated with the largest eigenvalue μ_n of $\mathbf{Y}/\sqrt{n}$.

B.B.P. phase transition

- if $\lambda \le 1$ $\begin{cases} \mu_n & \xrightarrow[n\to\infty]{\text{a.s.}} 2 \\ \mathbf{X}\cdot\hat{\mathbf{x}}_n & \xrightarrow[n\to\infty]{\text{a.s.}} 0 \end{cases}$
- if $\lambda > 1$ $\begin{cases} \mu_n & \xrightarrow[n\to\infty]{\text{a.s.}} \sqrt{\lambda} + \frac{1}{\sqrt{\lambda}} > 2 \\ |\mathbf{X}\cdot\hat{\mathbf{x}}_n| & \xrightarrow[n\to\infty]{\text{a.s.}} \sqrt{1-1/\lambda} > 0 \end{cases}$

Baik et al., 2005; Benaych-Georges and Nadakuditi, 2011

Questions

- PCA fails when $\lambda \leq 1$, but is it still possible to recover the signal?

Questions

- PCA fails when $\lambda \leq 1$, but is it still possible to recover the signal?
- When $\lambda > 1$, is PCA optimal?

Questions

- PCA fails when $\lambda \leq 1$, but is it still possible to recover the signal?
- When $\lambda > 1$, is PCA optimal?
- More generally, what is the best achievable estimation performance in both regimes?

MMSE and information-theoretic threshold

Definitions

“MMSE” $=$ Minimal Mean Square Error

$$\begin{aligned}\mathrm{MMSE}_n &= \min_{\hat{\theta}} \frac{1}{n^2}\mathbb{E}\left\|\mathbf{X}\mathbf{X}^\intercal - \hat{\theta}(\mathbf{Y})\right\|^2 \\ &= \frac{1}{n^2}\sum_{1\le i,j\le n}\left(X_iX_j - \mathbb{E}[X_iX_j|\mathbf{Y}]\right)^2 \le \underbrace{\mathbb{E}_{P_0}[X^2]^2}_{\text{Dummy MSE}}\end{aligned}$$

The information-theoretic threshold is the critical value λ_c such that

- if $\lambda > \lambda_c$, $\quad \lim_{n\to\infty} \mathrm{MMSE}_n <$ Dummy MSE
- if $\lambda < \lambda_c$, $\quad \lim_{n\to\infty} \mathrm{MMSE}_n =$ Dummy MSE

Related work

A short overview

- Approximate Message Passing (AMP) algorithms: Rangan and Fletcher, 2012, Deshpande and Montanari, 2014; Lesieur et al., 2015 allows to derive the MMSE when AMP is optimal.
- In presence of a "hard phase", Barbier et al., 2016 uses AMP and spatial coupling techniques to compute the MMSE under some additional assumptions.
- Banks et al., 2016; Perry et al., 2016 obtained bounds on the information-theoretic threshold by second moment computations and contiguity.

Main result

Limiting formula for the MMSE

Theorem

$$\mathrm{MMSE}_n \xrightarrow[n\to\infty]{} \underbrace{\mathbb{E}_{P_0}[X^2]^2}_{\text{Dummy MSE}} - q^*(\lambda)^2$$

where $q^*(\lambda)$ is the maximizer of

$$q \geq 0 \mapsto \mathbb{E}_{\substack{X_0\sim P_0 \\ Z_0\sim\mathcal{N}}}\left[\log \int_{x_0} dP_0(x_0) e^{\sqrt{\lambda q} Z_0 x_0 + \lambda q X_0 x_0 - \frac{\lambda q}{2} x_0^2}\right] - \frac{\lambda}{4} q^2$$

Proof ideas

A planted spin system

$\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \frac{1}{\mathcal{Z}_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})}$ where

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2 \,.$$

Proof ideas

A planted spin system

$$\mathbb{P}(\mathbf{X} = \mathbf{x} \,|\, \mathbf{Y}) = \frac{1}{\mathcal{Z}_n} P_0(\mathbf{x}) e^{H_n(\mathbf{x})} \text{ where}$$

$$H_n(\mathbf{x}) = \sum_{i<j} \sqrt{\frac{\lambda}{n}} Y_{i,j} x_i x_j - \frac{\lambda}{2n} x_i^2 x_j^2 \,.$$

Two step proof:

- Lower bound: Guerra's interpolation technique. Adapted in Korada and Macris, 2009; Krzakala et al., 2016.

$$\begin{cases} \mathbf{Y} & = \sqrt{t} \quad \sqrt{\lambda/n} \quad \mathbf{X}\mathbf{X}^\intercal \quad + \quad \mathbf{Z} \\ \mathbf{Y}' & = \sqrt{1-t} \quad \sqrt{\lambda} \quad \mathbf{X} \quad + \quad \mathbf{Z}' \end{cases}$$

- Upper bound: Cavity computations (Mézard et al., 1987).
  Aizenman-Sims-Starr scheme: Aizenman et al., 2003; Talagrand, 2010.

Some curves

Recall $\mathbf{Y} = \sqrt{\lambda/n}\mathbf{X}\mathbf{X}^{\intercal} + \mathbf{Z}$, where $(X_i)_{1 \leq i \leq n} \overset{\text{i.i.d.}}{\sim} P_0$.

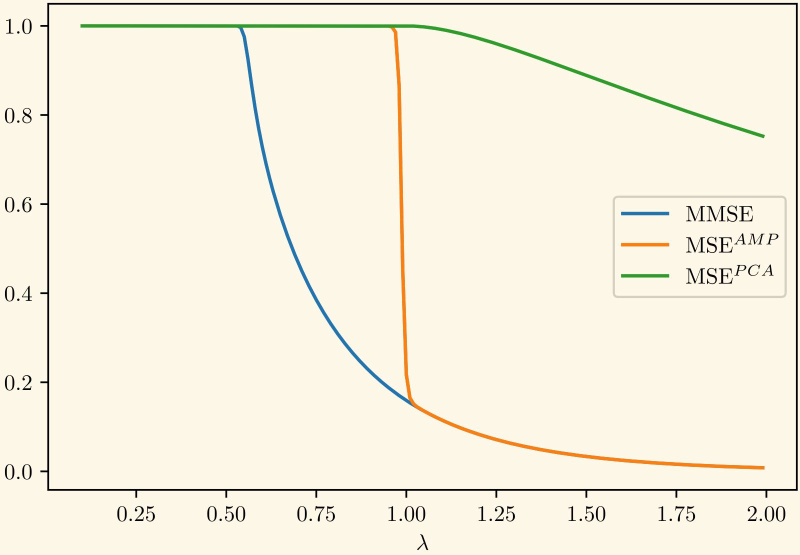
- We will plot the MMSE and $\mathrm{MSE}^{\mathrm{PCA}}$ curves for priors of the form

$$X_i = \begin{cases} \sqrt{\frac{1-p}{p}} & \text{with probability } p \\ -\sqrt{\frac{p}{1-p}} & \text{with probability } 1-p \end{cases}$$

  for some $p \in (0,1)$.

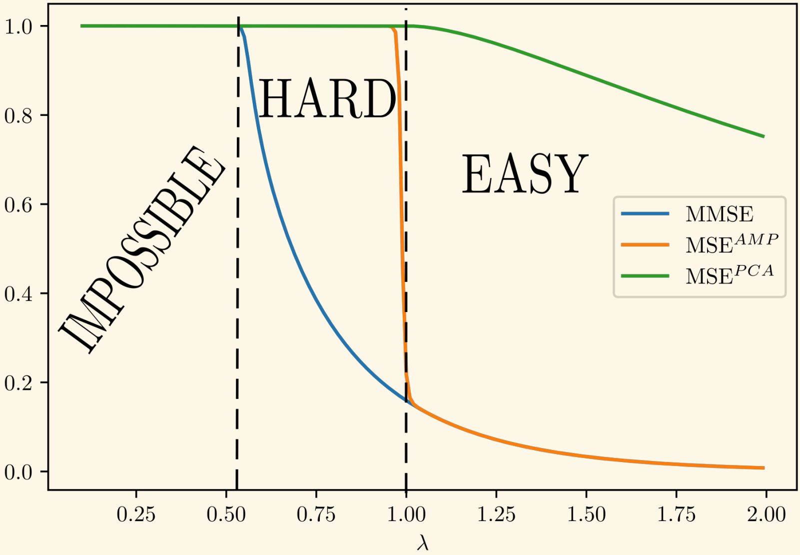
- One can show (similarly to Deshpande and Abbe, 2016) that the corresponding matrix estimation problem is, in some sense, equivalent to the community detection problem with 2 asymmetric communities of sizes pn and $(1-p)n$.

Plot of MMSE



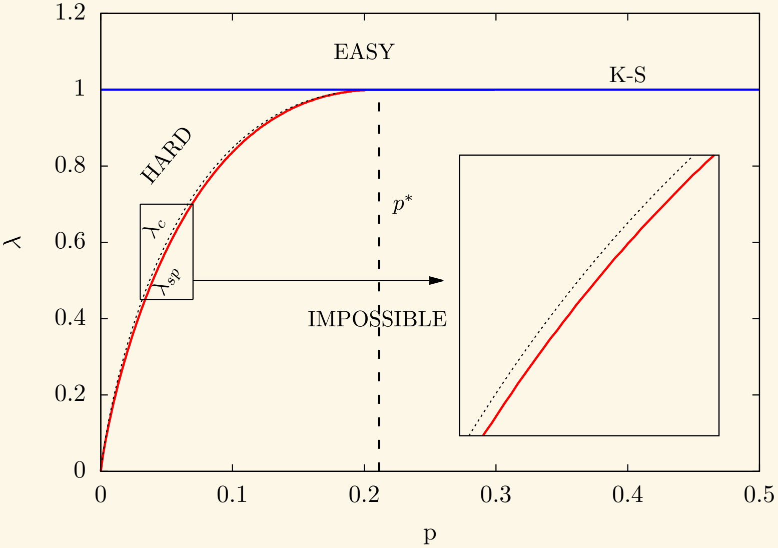
MMSE, $\mathrm{MSE}^{\mathrm{PCA}}$ and $\mathrm{MSE}^{\mathrm{AMP}}$, $p = 0.05$.

Plot of MMSE



MMSE, $\mathrm{MSE}^{\mathrm{PCA}}$ and $\mathrm{MSE}^{\mathrm{AMP}}$, $p = 0.05$.

Phase diagram for asymmetric community detection



Phase diagram from Caltagirone et al., 2017

Thank you for your attention.

Any questions?

References I

references.bib

- Aizenman, Michael, Robert Sims, and Shannon L Starr (2003). "Extended variational principle for the Sherrington-Kirkpatrick spin-glass model". In: *Physical Review B* 68.21, p. 214403.
- Baik, Jinho, Gérard Ben Arous, and Sandrine Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". In: *Annals of Probability*, pp. 1643–1697.
- Banks, Jess et al. (2016). "Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization". In: *arXiv preprint arXiv:1607.05222v2*.
- Barbier, Jean et al. (2016). "Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula". In: *Advances in Neural Information Processing Systems*, pp. 424–432.
- Benaych-Georges, Florent and Raj Rao Nadakuditi (2011). "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices". In: *Advances in Mathematics* 227.1, pp. 494–521.

References II

- Caltagirone, Francesco, Marc Lelarge, and Léo Miolane (2017). "Recovering asymmetric communities in the stochastic block model". In: *IEEE Transactions on Network Science and Engineering*.
- Deshpande, Yash and Emmanuel Abbe (2016). "Asymptotic mutual information for the balanced binary stochastic block model". In: *Information and Inference*, iaw017.
- Deshpande, Yash and Andrea Montanari (2014). "Information-theoretically optimal sparse PCA". In: *2014 IEEE International Symposium on Information Theory*. IEEE, pp. 2197–2201.
- Korada, Satish Babu and Nicolas Macris (2009). "Exact solution of the gauge symmetric p-spin glass model on a complete graph". In: *Journal of Statistical Physics* 136.2, pp. 205–230.
- Krzakala, Florent, Jiaming Xu, and Lenka Zdeborová (2016). "Mutual information in rank-one matrix estimation". In: *Information Theory Workshop (ITW), 2016 IEEE*. IEEE, pp. 71–75.

References III

- Lesieur, Thibault, Florent Krzakala, and Lenka Zdeborová (2015). "Phase transitions in sparse PCA". In: *IEEE International Symposium on Information Theory, ISIT 2015, Hong Kong, China, June 14-19, 2015*. IEEE, pp. 1635–1639. ISBN: 978-1-4673-7704-1. DOI: 10.1109/ISIT.2015.7282733. URL: http://dx.doi.org/10.1109/ISIT.2015.7282733.
- Massoulié, Laurent (2014). "Community detection thresholds and the weak Ramanujan property". In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM, pp. 694–703.
- Mézard, Marc, Giorgio Parisi, and Miguel Virasoro (1987). *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Publishing Co Inc.
- Mossel, Elchanan, Joe Neeman, and Allan Sly (2013). "A proof of the block model threshold conjecture". In: *arXiv preprint arXiv:1311.4115*.
- – (2015). "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* 162.3-4, pp. 431–461.
- Perry, Amelia et al. (2016). "Optimality and Sub-optimality of PCA for Spiked Random Matrices and Synchronization". In: *arXiv preprint arXiv:1609.05573*.

References IV

- Rangan, Sundeep and Alyson K Fletcher (2012). "Iterative estimation of constrained rank-one matrices in noise". In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, pp. 1246–1250.
- Talagrand, Michel (2010). *Mean field models for spin glasses: Volume I: Basic examples*. Vol. 54. Springer Science & Business Media.