Semidefinite Programs for Exact Recovery of a Hidden Community (and Many Communities)

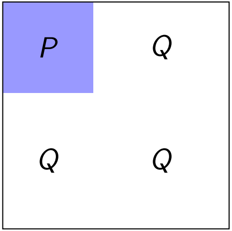
Bruce Hajek [1] Yihong Wu [1] Jiaming Xu [2]

[1]University of Illinois at Urbana-Champaign

[2]Simons Insitute, UC Berkeley

June 23, 2016

Hidden community model [Deshpande-Montanari '13]



- Data: $n \times n$ symmetric matrix A with empty diagonal
- Community $C^* \subset [n]$ of size K uniform at random, such that

$$A_{ij} \sim \begin{cases} P & \text{both } i \text{ and } j \in C \\ Q & \text{otherwise} \end{cases}$$

- (K, P, Q) varies with n
- Goal: exact recovery of C from A

$$\mathbb{P}\{\widehat{C} = C^*\} \xrightarrow{n \to \infty} 1$$

- Fruitful venue for stuying computational aspects of statistical problems

Examples

Planted dense subgraph

$$P = \mathrm{Bern}(p), Q = \mathrm{Bern}(q), \quad p > q$$

- $A =$ adjancency matrix of $G(n, q)$ planted with $G(K, p)$
- [Alon et al '98, McSherry '01, Arias-Castro-Verzelen '14, Chen-Xu 14, Montanari '15, ...]

Examples

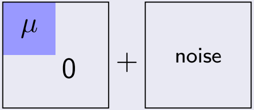
Planted dense subgraph

$$P = \mathrm{Bern}(p), Q = \mathrm{Bern}(q), \quad p > q$$

- $A =$ adjancency matrix of $G(n, q)$ planted with $G(K, p)$
- [Alon et al '98, McSherry '01, Arias-Castro-Verzelen '14, Chen-Xu 14, Montanari '15, ...]

Submatrix localization

$$P = \mathcal{N}(0, \mu), Q = \mathcal{N}(0, 1), \quad \mu > 0$$

- $A =$ 

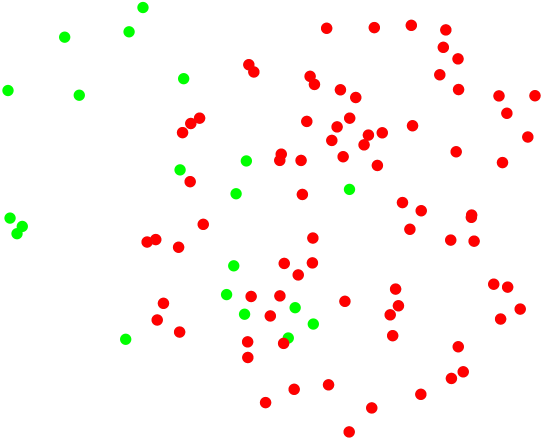
- [Shabalin et al '09, Butucea-Ingster '11, Kolar et al '11, Ma-W '13, Cai et al '15, ...]

Running example:
Plated Dense Subgraph

Planted dense subgraph – graph view

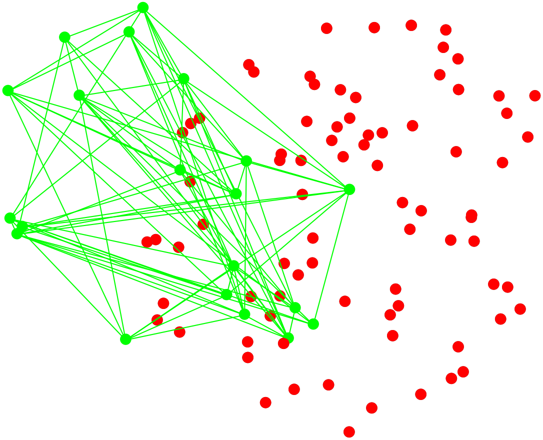
Planted dense subgraph – graph view

1. A community of K vertices are chosen randomly



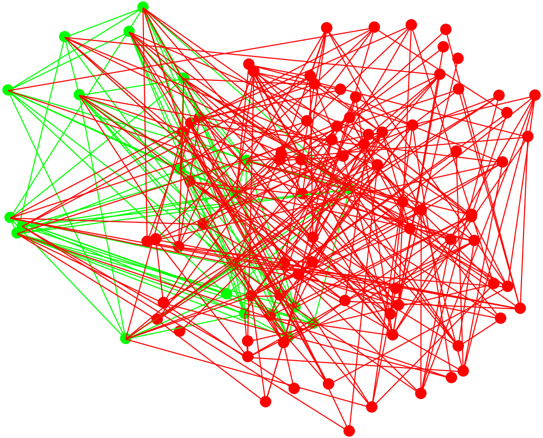
Planted dense subgraph – graph view

1. A community of K vertices are chosen randomly
2. For every pair of nodes in the community, add an edge w.p. p



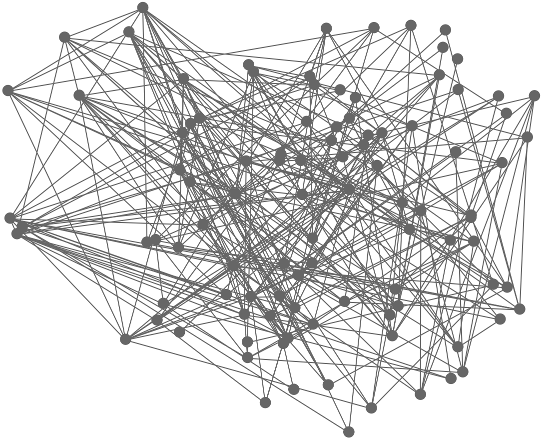
Planted dense subgraph – graph view

1. A community of K vertices are chosen randomly
2. For every pair of nodes in the community, add an edge w.p. p
3. For other pairs of nodes, add an edge w.p. q

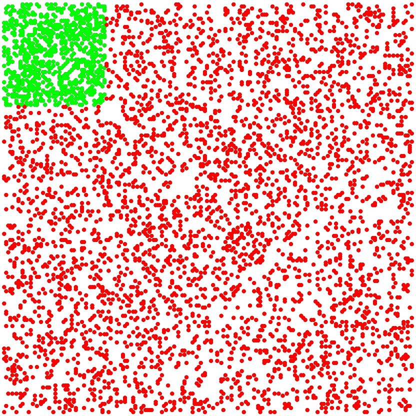


Planted dense subgraph – graph view

1. A community of K vertices are chosen randomly
2. For every pair of nodes in the community, add an edge w.p. p
3. For other pairs of nodes, add an edge w.p. q

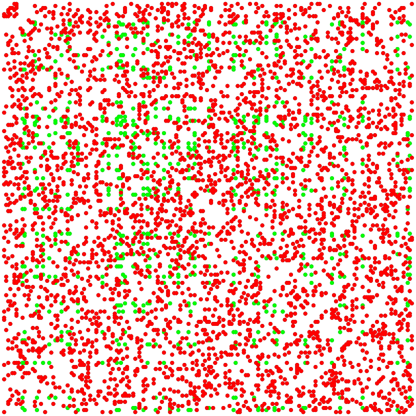


Planted dense subgraph – adjacency matrix view



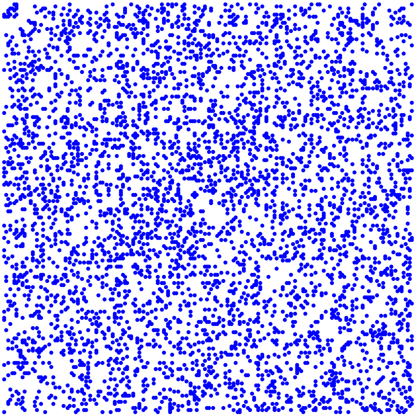
$n = 200,\ K = 50,\ p = 0.3,\ q = 0.1$

Planted dense subgraph – adjacency matrix view



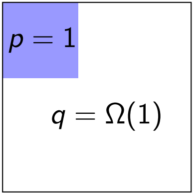
$n = 200,\ K = 50,\ p = 0.3,\ q = 0.1$

Planted dense subgraph – adjacency matrix view



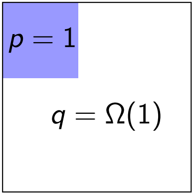
$n = 200,\ K = 50,\ p = 0.3,\ q = 0.1$

Computational gap in planted Clique



- $K = \Omega(\log n)$: exact recovery is possible via maximum likelihood
- $K = \Omega(\sqrt{n})$: exact recovery is attainable in poly-time [Alon et al. '98]
- $K = o(\sqrt{n})$: exact recovery is believed to be hard [Deshpande-Montanari '15] [Meka-Potechin-Wigderson '15], ...

Computational gap in planted Clique



- $K = \Omega(\log n)$: exact recovery is possible via maximum likelihood
- $K = \Omega(\sqrt{n})$: exact recovery is attainable in poly-time [Alon et al. '98]
- $K = o(\sqrt{n})$: exact recovery is believed to be hard [Deshpande-Montanari '15] [Meka-Potechin-Wigderson '15], ...

What about dense subgraphs ~~clique~~?

Linear community size

- $K = \rho\, n$
- $p = \frac{a \log n}{n}$ and $q = \frac{b \log n}{n}$

Theorem (Hajek-W-Xu Trans. IT 16)

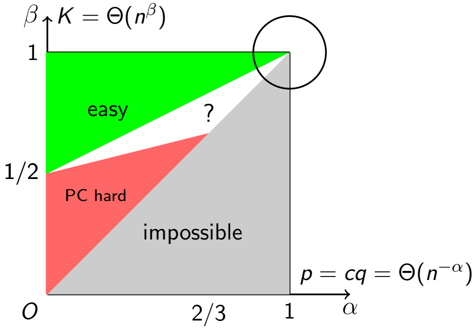
- *If $\rho > \rho^*$, exact recovery is possible in polynomial-time.*
- *If $\rho < \rho^*$, exact recovery is impossible.*

Remarks

- $\rho^* = 1/(a - \tau^* \log \frac{\mathrm{e}a}{\tau^*})$ with $\tau^* = \frac{a-b}{\log a - \log b}$
- Convex (SDP) relaxation works

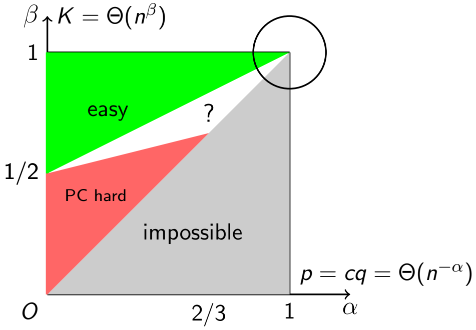
Sublinear community size

[Hajek-W-Xu, COLT '15]



- $K = \Omega(n)$: SDP works
- $K = n^{1-\epsilon}$: no known poly-time algorithm
- Where is the SDP barrier?

[Hajek-W-Xu, COLT '15]



- $K = \Omega(n)$: SDP works
- $K = n^{1-\epsilon}$: no known poly-time algorithm
- Where is the SDP barrier? $K = \Theta(\frac{n}{\log n})$

SDP Relaxation vs. Information-Theoretic Limits

Main results: For both planted dense subgraph (Bernoulli) and submatrix localization (Gaussian)

- $K = \omega(\frac{n}{\log n})$: SDP attains the info-theoretic limit with sharp constants
- $K = \Theta(\frac{n}{\log n})$: SDP is order-wise optimal, but strictly suboptimal by a constant factor
- $K = o(\frac{n}{\log n})$ and $K \to \infty$: SDP is order-wise suboptimal

SDP Relaxation vs. Information-Theoretic Limits

- Log-likelihood ratio matrix L

$$L_{ij} = \log \frac{dP}{dQ}(A_{ij}), i \neq j, \quad L_{ii} = 0$$

- Let $\xi =$ indicator of C.

Maximum likelihood estimator $=$ find densest K-subgraph

$$\begin{aligned} \hat{\xi}_{\mathrm{MLE}} = \arg\max_{\xi} & \sum_{i,j} L_{ij} \xi_i \xi_j \\ \text{s.t.} \quad & \xi \in \{0,1\}^n \\ & \langle \xi, \mathbf{1} \rangle = K. \end{aligned}$$

Lift: $Z = \xi\xi^*$

$$
\begin{aligned}
\hat{Z}_{\mathrm{MLE}} = \arg\max_{Z} \; & \langle L, Z \rangle \\
\text{s.t.} \quad & \mathrm{rank}(Z) = 1 \\
& Z_{ii} \le 1 \quad \forall i \in [n] \\
& Z_{ij} \ge 0, \quad \forall i, j \in [n] \\
& \langle \mathbf{I}, Z \rangle = K \\
& \langle \mathbf{J}, Z \rangle = K^2
\end{aligned}
$$

Semidefinite programming

Natural SDP relaxation:

$$
\begin{aligned}
\hat{Z}_{\mathrm{SDP}} = \arg\max_{Z} \ & \langle L, Z \rangle \\
\text{s.t.} \ & Z \succeq 0 \\
& Z_{ii} \le 1 \quad \forall i \in [n] \\
& Z \ge 0 \\
& \langle \mathbf{I}, Z \rangle = K \\
& \langle \mathbf{J}, Z \rangle = K^2
\end{aligned}
$$

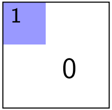
Semidefinite programming

Natural SDP relaxation:

$$
\begin{aligned}
\hat{Z}_{\mathrm{SDP}} = \arg\max_{Z} \ & \langle L, Z \rangle \\
\text{s.t.} \ & {\color{red} Z \succeq 0} \\
& Z_{ii} \leq 1 \quad \forall i \in [n] \\
& Z \geq 0 \\
& \langle \mathbf{I}, Z \rangle = K \\
& \langle \mathbf{J}, Z \rangle = K^2
\end{aligned}
$$

Goal:

$$\mathbb{P}\left\{ \widehat{Z}_{\mathrm{SDP}} = \widehat{Z}_{\mathrm{MLE}} = \right.$$



$$\left. \right\} \to 1$$

Define

$$e(i, C^*) = \sum_{j \in C^*} L_{ij}, \quad i \in [n], \quad , \beta = -D(Q \| P).$$

Analysis of SDP

Theorem

- *Sufficient condition*: $\widehat{Z}_{\rm SDP} = Z^*$, *if*

$$\min_{i \in C^*} e(i, C^*) - \max\left\{\max_{j \notin C^*} e(j, C^*), K\beta\right\} > \|L - \mathbb{E}[L]\| - \beta$$

Analysis of SDP

Theorem

- *Sufficient condition:* $\widehat{Z}_{\rm SDP} = Z^*$, *if*

$$\min_{i \in C^*} e(i, C^*) - \max\left\{\max_{j \notin C^*} e(j, C^*), K\beta\right\} > \|L - \mathbb{E}[L]\| - \beta$$

- *Necessary condition: If* $Z^* \in \widehat{Z}_{\rm SDP}$, *then*

$$\min_{i \in C^*} e(i, C^*) - \max_{j \notin C^*} e(j, C^*) \geq \sup_{1 \leq a \leq K}\left\{V(a) - \frac{a}{K}\max_{j \notin C^*} e(j, C^*)\right\},$$

where

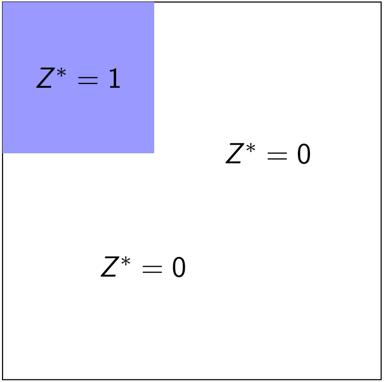
- $V(a) = \max\{\langle L_{\overline{C^*} \times \overline{C^*}}, Z\rangle : Z \succeq 0, Z \geq 0, \mathrm{Tr}(Z) = 1, \langle \mathbf{J}, Z\rangle = a\}$ *is the value of an (simpler) auxilliary SDP*

Remarks

- To apply this result, $\min, \max, \|L - \mathbb{E}[L]\|$, etc concentrate
- Sufficient condition proof: construction of dual witnesses (standard)

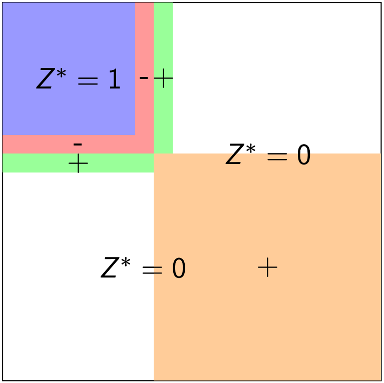
Proof of necessary condition

- Primal proof: random perturbation of the ground truth to establish integrality gap



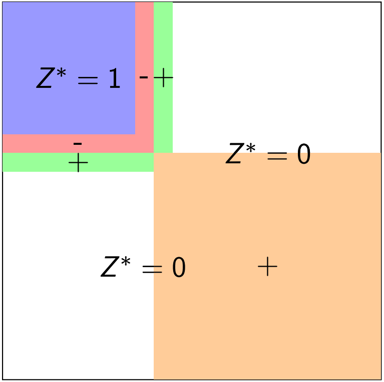
Proof of necessary condition

- Primal proof: random perturbation of the ground truth to establish integrality gap



Proof of necessary condition

- Primal proof: random perturbation of the ground truth to establish integrality gap



- Dual proof: non-existence of dual witness

Multiple communities

k communities: MLE $\Rightarrow$ SDP relaxation

SBM with k communities and parameter (p, q)

$$\max \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_\ell \boldsymbol{\theta}_\ell^\top \rangle$$

$$\begin{aligned} \text{s.t. } \quad & \boldsymbol{\theta}_\ell \in \{0,1\}^n \\ & \langle \boldsymbol{\theta}_\ell, \mathbf{1} \rangle = n/k \\ & \langle \boldsymbol{\theta}_\ell, \boldsymbol{\theta}_{\ell'} \rangle = 0, \ell \neq \ell' \end{aligned}$$

k communities: MLE $\Rightarrow$ SDP relaxation

SBM with k communities and parameter (p, q)

$$\begin{aligned} \max \quad & \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_\ell \boldsymbol{\theta}_\ell^\top \rangle \\ \text{s.t.} \quad & \boldsymbol{\theta}_\ell \in \{0,1\}^n \\ & \langle \boldsymbol{\theta}_\ell, \mathbf{1} \rangle = n/k \\ & \langle \boldsymbol{\theta}_\ell, \boldsymbol{\theta}_{\ell'} \rangle = 0, \ell \neq \ell' \end{aligned}
\quad \xLeftrightarrow{\text{lift: } Z = \sum_{\ell=1}^{k} \boldsymbol{\theta}_\ell \boldsymbol{\theta}_\ell^\top} \quad
\begin{aligned} \max \quad & \langle A, Z \rangle \\ \text{s.t.} \quad & \mathrm{rank}(Z) = k \\ & Z_{ii} = 1 \quad \forall i \in [n] \\ & Z_{ij} \geq 0, \quad \sum_j Z_{ij} = n/k \end{aligned}$$

SBM with k communities and parameter (p, q)

$$
\begin{aligned}
\max \ & \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_{\ell} \boldsymbol{\theta}_{\ell}^{\top} \rangle \\
\text{s.t.} \ & \boldsymbol{\theta}_{\ell} \in \{0,1\}^n \\
& \langle \boldsymbol{\theta}_{\ell}, \mathbf{1} \rangle = n/k \\
& \langle \boldsymbol{\theta}_{\ell}, \boldsymbol{\theta}_{\ell'} \rangle = 0, \ell \neq \ell'
\end{aligned}
\quad
\xLeftrightarrow{\text{lift: } Z = \sum_{\ell=1}^{k} \boldsymbol{\theta}_{\ell} \boldsymbol{\theta}_{\ell}^{\top}}
\quad
\begin{aligned}
\max \ & \langle A, Z \rangle \\
\text{s.t.} \ & Z \succeq 0 \\
& Z_{ii} = 1 \quad \forall i \in [n] \\
& Z_{ij} \geq 0, \quad \sum_{j} Z_{ij} = n/k
\end{aligned}
$$

k communities: MLE $\Rightarrow$ SDP relaxation

SBM with k communities and parameter (p, q)

$$\begin{aligned} \max \ & \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^\top \rangle \\ \text{s.t.} \ & \boldsymbol{\theta}_{\boldsymbol{\ell}} \in \{0,1\}^n \\ & \langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \mathbf{1} \rangle = n/k \\ & \langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \boldsymbol{\theta}_{\boldsymbol{\ell'}} \rangle = 0, \ell \neq \ell' \end{aligned} \quad \xLeftrightarrow{\text{lift: } Z = \sum_{\ell=1}^{k} \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^\top} \quad \begin{aligned} \max \ & \langle A, Z \rangle \\ \text{s.t.} \ & Z \succeq 0 \\ & Z_{ii} = 1 \quad \forall i \in [n] \\ & Z_{ij} \geq 0, \quad \sum_j Z_{ij} = n/k \end{aligned}$$

$$\text{Goal: } \mathbb{P}\left\{ \widehat{Z}_{\mathrm{SDP}} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix} \right\} \to 1$$

k equal-sized communities: optimal recovery via SDP

Theorem (Hajek-W-Xu '15)

For a fixed k communities with $p = a \log n/n$ and $q = b \log n/n$.

- *If $\sqrt{a} - \sqrt{b} > \sqrt{k}$, exact recovery is attained via SDP in poly-time.*
- *If $\sqrt{a} - \sqrt{b} < \sqrt{k}$, exact recovery is impossible.*

k equal-sized communities: optimal recovery via SDP

Theorem (Hajek-W-Xu '15)

For a fixed k communities with $p = a \log n/n$ and $q = b \log n/n$.

- *If $\sqrt{a} - \sqrt{b} > \sqrt{k}$, exact recovery is attained via SDP in poly-time.*
- *If $\sqrt{a} - \sqrt{b} < \sqrt{k}$, exact recovery is impossible.*

Remarks

- Extended to $k = o(\log n)$ in [Agarwal-Bandeira-Koiliaris-Kolla '15]

k equal-sized communities: optimal recovery via SDP

Theorem (Hajek-W-Xu '15)

For a fixed k communities with $p = a \log n/n$ and $q = b \log n/n$.

- *If $\sqrt{a} - \sqrt{b} > \sqrt{k}$, exact recovery is attained via SDP in poly-time.*
- *If $\sqrt{a} - \sqrt{b} < \sqrt{k}$, exact recovery is impossible.*

Remarks

- Extended to $k = o(\log n)$ in [Agarwal-Bandeira-Koiliaris-Kolla '15]
- Extended to the case with multiple unequal-sized clusters [Perry-Wein '15]

When does SDP cease to be optimal?

Theorem (Hajek-W.-Xu '16)

- $k \ll \log n$: *SDP achieves the optimal exact recovery threshold.*
- $k \geq c \log n$: *SDP is suboptimal by a constant factor.*
- $k \gg \log n$: *SDP is order-suboptimal.*

Remarks

- A "hard but informationally possible" regime is conjectured to exist for exact recovery when $k \gg \log n$ [Chen-Xu '14]

Some remaining problems

- Can the computational gap for exact recovery be bridged by any polynomial time algorithm? (SoS hardness result or reduction to PC would offer further evidence for “no” answer.)
- Approximate recovery? (Current proof only rules out exact recovery.)

Some remaining problems

- Can the computational gap for exact recovery be bridged by any polynomial time algorithm? (SoS hardness result or reduction to PC would offer further evidence for "no" answer.)
- Approximate recovery? (Current proof only rules out exact recovery.)

Thank you!

Necessary condition for optimality of SDP

EXTRA SLIDES NOT INCLUDED IN ORIGINAL Let $M = L_{(C^*)^c \times (C^*)^c}$ denote the submatrix of L outside the community. For $a \in \mathbb{R}$, consider the (random) value of the following SDP:

$$\begin{aligned} V(a) \triangleq \max_{Z} \; & \langle M, Z \rangle \\ \text{s.t.} \; & Z \succeq 0 \\ & Z \geq 0 \\ & \mathrm{Tr}(Z) = 1 \\ & \langle \mathbf{J}, Z \rangle = a. \end{aligned} \tag{1}$$

Necessary condition for optimality of SDP

Theorem (Necessary condition for SDP)

If $Z^* \in \widehat{Z}_{\rm SDP}$*, then*

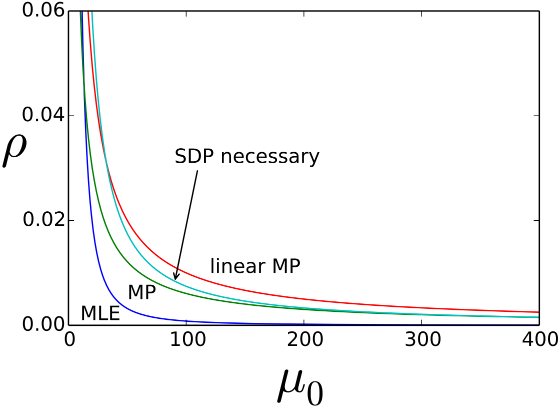
$$\min_{i \in C^*} e(i, C^*) - \max_{j \notin C^*} e(j, C^*) \geq \sup_{1 \leq a \leq K} \left\{ V(a) - \frac{a}{K} \max_{j \notin C^*} e(j, C^*) \right\}. \quad (2)$$

Weaker necessary condition (set $a = K$):

$$\min_{i \in C^*} e(i, C^*) \geq V(K)$$

.

SDP vs. MLE, message passing, and linear MP



Phase diagram for the Gaussian model
with $K = \rho n / \log n$ and $\mu = \mu_0 \log n / \sqrt{n}$.