Lecture 9  Detection threshold for SBM

- Planted partition model and overview
  - Motivation: Community detection.
  - Given latent community labels $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in \{\pm 1\}^n$, generate a random graph with weighted adjacency matrix $A=(A_{ij})$,
    $A_{ij} \sim \begin{cases} P & \text{if } \sigma_i = \sigma_j \\ Q & \text{if } \sigma_i \neq \sigma_j \end{cases}$
  - Recovery goal: Given $A$, recover the labels $\sigma$ accurately.

- Two prominent special cases

  1. Stochastic Block model (SBM):

     Here $P = \text{Bern}(p)$ and $Q = \text{Ber}(q)$

     Here $V_+ = \{ i : \sigma_i = + \}$ and $V_- = \{ i : \sigma_i = - \}$

     are two communities.

     $A$ is the adjacency matrix, where w.p. $p$ if $i,j$ in same community

     and w.p. $q$ if $i,j$ in two different communities

     $P > q$: "assortative": densely connected inside, like friendship network

     $P < q$: "disassortative": sparsely connected inside.
Like predator-prey network.

The community structure is determined by $\Sigma$.

Two special cases to generate $\Sigma$:

- iid model: each $\Sigma_i \sim \text{iid Rad}(1/2)$

- Exact bisection model: $|V+| = |V_-| = \frac{n}{2}$ (assumed even)
  
  and the partition is chosen uniformly at random from all bisections.

  The two models behave very similarly.

\( \Rightarrow \) spiked Wigner model (Rank-1 deformation)

Here $P = N(\sqrt{\frac{A}{n}}, 1)$ and $Q = N(-\sqrt{\frac{1}{n}}, 1)$

In matrix notation:

$$A = \sqrt{\frac{A}{n}} \Sigma \Sigma^T + Z$$

Where $Z$ is such that $\{Z_{ij} : 1 \leq i < j \leq n \}$ are iid $N(0,1)$.

$A$ can be viewed as rank-1 perturbation of Gaussian Wigner matrix $Z$.

Targeted Results:

As opposed to the treatment of the planted clique problem, we will focus on

- Sharp threshold, i.e., find the exact constant in the function
limit (and achieve them with fast alg.)

- "Sparse" graphs, where the edge density $p, q$ tends to 0.

Unlike the planted clique model $G(n, \frac{1}{2}, k)$.

Recovery guarantees.

1) Detection: Here there is a null model. For example,

- For spiked Wigner model, the null hypothesis

  $H_0: A_{ij} \sim N(0,1)$$
  H_1: A = \frac{\sqrt{\lambda}}{\sqrt{n}} \Theta^T + Z$

  The sharp threshold is given by $\lambda = 1$: for any fixed $C > 0$.

  If $\lambda > 1 + 2$, then it is possible to test two hypothesis with vanishing error prob.

  If $\lambda \leq 1 - \varepsilon$, then it is impossible to do so.

- For SBM, the null hypothesis

  $H_0: A \sim G(n, \frac{\Phi}{2})$
  $H_1: A \sim \text{SBM}(n, p, q)$

  The most intense case is $p = \frac{\alpha}{n} \text{ and } q = \frac{\beta}{n}$ for constants $\alpha, \beta$.

  The sharp threshold is $\frac{(\alpha - \beta)^2}{2(\alpha + \beta)} > 1$.

2) Correlated (Weak) Recovery

Here and below, there is no null model. The goal is to recover $\Theta$ better than random guess.\]
Let $\hat{\theta} = \hat{\theta}(A)$ be the estimator. Its overlap with true label $\theta$ is

$$|\langle \hat{\theta}, \theta \rangle|$$

and the number of misclassification errors (up to a global sign flip) is

$$\ell(\theta, \hat{\theta}) \triangleq \min_{\sigma \in \{1, -1\}^{n}} \| \hat{\theta} + \sigma \theta \|_1 = n - |\langle \hat{\theta}, \theta \rangle|$$

In iid setting, random guess would yield by CLT

$$|\langle \hat{\theta}, \theta \rangle| = O(\sqrt{n}) \quad \text{w.h.p}$$

and $\mathbb{E}[|\langle \hat{\theta}, \theta \rangle|] = o(n)$

The goal of correlated recovery is to achieve a strictly positive correlation, namely

$$\mathbb{E}[|\langle \hat{\theta}, \theta \rangle|] = \sqrt{n}.$$  

Remark: In general, detection and estimation are two different problems. But for both Sparse SBM and Spiked Wigner, the thresholds coincide.

3. (Almost) exact recovery: Almost exact recovery means achieving a vanishing misclassification rate, i.e., $\mathbb{E}\left[\ell(\hat{\theta}, \theta)\right] = o(n)$.

Typically, the threshold is given by $H^2(\theta, \hat{\theta}) \gg \frac{1}{n}$, where

$$H^2(\theta, \hat{\theta}) \triangleq \mathbb{E}_{\theta} \left[ (\sqrt{\frac{p}{\theta}} - 1)^2 \right] = \int (\sqrt{tp} - t\theta)^2$$

Squared Hellinger
Exact Decay means $\ell(0,\delta) = 0$ w.p. convergent to 1.

Typically, the exact decay threshold is given by

$$H^2(p,q) = \frac{2 \log n}{n}$$

- Preliminary on Binary Hypothesis Test:

Consider a binary hypothesis test:

$H_0: X \sim \mathcal{O}$

$H_1: X \sim P$

A test $\phi: X \rightarrow \{0,1\}$, where $\phi(x) = 0$ means "decy on $H_0$ is not" $\phi(x) = 1$ means "decy on $H_1$ is not."

Type I error = $\mathcal{O}(\phi(x)=0)$ "false positive"

Type II error = $P(\phi(x)=1)$ "false negative"

**Thm:** The minimum of Type I + Type II errors is 1 - TV$(p,\delta)$, where $TV(p,\delta) = \frac{1}{2} \sup_E \left[ |P - \delta| \right] = \frac{1}{2} \sup_{E} |P - \mathcal{O}|$

i.e.,

$$\min_{\phi} \mathbb{P}(\phi(x) = 0) + \mathbb{Q}(\phi(x) = 1) = 1 - TV(p,\delta)$$

**Pf:** Exercise. Hint: let $E = \{ \phi(x) = 1 \}$.

Remark: $TV(p,\delta) = 1 \iff P \perp \mathcal{Q}$, minimize Type I + II errors = 0
\[ \text{Def: } \chi^2 \text{-divergence between } P \text{ and } Q: \]
\[ \chi^2(P||Q) = \mathbb{E}_Q \left( \frac{P}{Q} - 1 \right)^2 = \int \frac{P^2}{Q} - 1 \]

Remark: \[ \chi^2(P||Q) \neq \chi^2(Q||P) \]

- \[ 2TV(P,Q) = \mathbb{E}_Q \left[ 1 - \frac{P}{Q} \right] \]
- \[ \leq \sqrt{\mathbb{E}_Q \left[ 1 - \frac{P}{Q} \right]^2} \]
- \[ = \sqrt{\chi^2(P||Q)} \]

Thus, if \[ \chi^2(P||Q) = o(1) \], then \[ TV(P,Q) = o(1) \]

Def [contiguity]: We say \( (P_n) \) is contiguous to \( (Q_n) \), if for any sequence of events \( E_n, Q_n \cap E_n \to 0 \Rightarrow P_n(E_n) \to 0 \)

Remark: Contiguity \[ \Rightarrow \min \sum \text{of Type I+II errors } Z \leq o(1) \], i.e., \[ TV(P,Q) \leq 1 - o(1) \]

For any test \( \phi \),
\[ \text{if } Q \phi(x) = 13 \to 0, \text{ then by contiguity} \]
Lemma 1 If $X^2(\mathbb{H}1/\mathbb{Q}) = O(1)$, then $P_n$ is contiguous to $\mathbb{Q}n$.

Proof: $P_n(\mathbb{E}n) = \mathbb{E}n \left[ \frac{P_n}{\mathbb{Q}n} \right ]$

\[
\leq \sqrt{\mathbb{E}n \left[ \left( \frac{P_n}{\mathbb{Q}n} \right )^2 \right ] - \mathbb{E}n \left[ 1_{\mathbb{Q}n}^2 \mathbb{E}n \right ]}
\]

\[
= \sqrt{\mathbb{E}n \left[ (\mathbb{Q}n)^2 + 1 \mathbb{Q}n(\mathbb{E}n) \right ]}
\]

$= O(1)$, if $\mathbb{Q}n(\mathbb{E}n) = O(1)$.

In many planted problems, $P_n$ is a mixture distribution, $\mathbb{Q}n$ is a simple distribution.

The following lemma will be useful for coping $\chi^2(n)$ mixture distribution, $\mathbb{Q}$ simple distribution.

Lemma 2 (Second-moment trick).

Suppose we have a parameter family of distributions $P_\theta: \theta \in \Theta$. Given a prior $\pi(\theta)$ parameterspace $\Theta$, 
Define the mixture distribution

\[ P_\pi \equiv \int P_\theta \cdot \pi(d\theta) \]

Then we have.

\[ \chi^2(P_\pi \| \pi) = \mathbb{E}[G(\theta, \tilde{\theta})] - 1, \text{ where} \]

\[ \theta, \tilde{\theta} \overset{\text{iid}}{\sim} \pi \text{ and } G(\theta, \tilde{\theta}) = \int \frac{P_\theta \cdot P_{\tilde{\theta}}}{Q} \]

**Proof:** Just Fubini's theorem

\[ \chi^2(P_\pi \| \pi) + 1 = \int \frac{P_\pi^2}{Q} \]

\[ = \int \frac{(P_\theta \pi(d\theta))^2}{Q} \]

\[ = \int \frac{\int P_\theta \pi(d\theta) \cdot \int P_{\tilde{\theta}} \pi(d\tilde{\theta})}{Q} \]

\[ = \int \pi(d\theta) \pi(d\tilde{\theta}) \cdot \int \frac{P_\theta \cdot P_{\tilde{\theta}}}{Q} \]

\[ = \mathbb{E}[G(\theta, \tilde{\theta})] \]

**Example (Gaussian):** Consider \( P_\theta = N(\theta, I_d) \) and \( Q = N(0, I_d) \). Let \( \pi \) be some distribution on \( \mathbb{R}^d \). Then

\[ \chi^2(P_\pi \| \pi) = \mathbb{E}[e^{\langle \theta, \tilde{\theta}\rangle}] - 1, \text{ where } \theta, \tilde{\theta} \overset{\text{iid}}{\sim} \pi \]
\[ G(\theta, \tilde{\theta}) = \int \frac{P_{0} P_{\tilde{\theta}}}{\theta} = E_{\theta} \left[ \frac{\int P_{0} P_{\tilde{\theta}}}{\theta^{2}} \right] \]

\[ = E_{\theta} \left[ e^{-\frac{||x-\theta||^{2}}{2} - \frac{||x-\tilde{\theta}||^{2}}{2} + \frac{1}{2} ||\theta + \tilde{\theta}||^{2}} \right] \]

\[ = E_{\theta} \left[ e^{-\frac{||x-\theta||^{2}}{2} + \langle x, \theta + \tilde{\theta} \rangle} \right] \]

\[ = e^{\frac{1}{2} ||\theta + \tilde{\theta}||^{2}} \cdot e^{-\frac{1}{2} ||\theta + \tilde{\theta}||^{2}} \]

\[ MGF \text{ of Gaussian} \]

\[ = e^{\langle \theta, \tilde{\theta} \rangle}. \]

**Detection threshold for SBM:**

Consider binary Hypothesis test:

\[ H_{0}: \quad G \sim G(n, \frac{p+q}{2}) \]

\[ H_{1}: \quad G \sim SBM(n, p, q) \]

Under SBM model, we assume the labels \( \sigma = (\sigma_{1}, \ldots, \sigma_{n}) \) are either:

- iid \( \text{Rad}(\frac{1}{2}) \) or drawn uniformly at random from \( \text{all bisections} \).

Assume \( p = \frac{a}{n} \) and \( q = \frac{b}{n} \) for two constants \( a, b \)

**Thm 1.** If \( \frac{(a-b)^{2}}{2(a+b)} > 1 \), detection is possible, i.e.,
\[TV(\text{Law}(G|H_0), \text{Law}(G|H_1)) = 1 + o(1)\].

If \(\frac{(a-b)^2}{2(ab)} < 1\), detection is impossible, i.e.,

\[TV(\text{Law}(G|H_0), \text{Law}(G|H_1)) \leq 1 - o(1)\].

By Lemma 7, it suffices to show \(X^2(\text{Law}(G|H_1)||\text{Law}(G|H_0)) = O(1)\).

The proof can be carried out in a very general setting. Consider \(p\) and \(q\) in place of \(\text{Ber}(p)\) and \(\text{Ber}(q)\).

Given each label \(\sigma \in \{\pm 3\}^n\),

\[P_0 \equiv \text{Law}(A|\sigma) = \prod_{i,j} (p 1_{\sigma_i = \sigma_j} + q 1_{\sigma_i = \sigma_j})\]

\[= \prod_{i,j} \left( \frac{pt\sigma}{2} + \frac{p-q}{2} \sigma_i \sigma_j \right)\]

The null distribution

\[P_0 \equiv \text{Law}(A|H_0) = \prod_{i,j} \left( \frac{pt\sigma}{2} \right)\]

Fix any two label vectors \(\sigma, \delta \in \{\pm 3\}^n\). Then

\[G(\sigma, \delta) = \int \frac{P_0 P_{\sigma}}{P_0} = \int \prod_{i,j} \left( \frac{pt\sigma}{2} + \frac{p-q}{2} \sigma_i \sigma_j \right) \frac{pt\sigma}{2}\]
\[ \begin{align*}
&= \frac{\Pi}{i \omega} \int \left( \frac{p}{2} + \frac{p}{2} \sigma_i \sigma_j + \frac{p}{2} \sigma_i \sigma_j \right) \\
&= \frac{\Pi}{i \omega} \left( 1 + \frac{p}{2} \sigma_i \sigma_j \right), \text{ where } \frac{\Pi}{i \omega} = \frac{p}{2} \sigma_i \sigma_j \\
&\leq \exp \left( \frac{p}{2} \sigma_i \sigma_j \right) \\
&\leq \exp \left( \frac{p}{2} \sum_{i,j} \sigma_i \sigma_j \right) \\
&= \exp \left( \frac{p}{2} \langle \sigma, \sigma \rangle^2 \right)
\end{align*} \]

By Lemma 2, we have:

\[ \chi^2 \left( \text{Law}(G | H_l) \mid \text{Law}(G | H_0) \right) + 1 \]

\[ \leq H_{\sigma, \sigma} \left[ e^{\frac{p}{2} \langle \sigma, \sigma \rangle^2} \right] \]

For SBM \((n, p, q)\).

\[ p = \int \frac{\Pi}{i \omega} \frac{(p-\sigma)^2}{2(p+\omega)} \]

\[ = \frac{p-q}{2(p+\omega)} + \frac{(1-p)(1-q)}{2(p+\omega)} \]

\[ = \frac{(a-b)^2}{2(ab)} + o \left( \frac{1}{n^2} \right) = \frac{1}{n} \left( 2 + o(1) \right), \]

where \( \frac{p-q}{2(ab)} \).
Now consider two cases.

- IID Rad(½) labels: Suppose σ₀, σ̄; iid Rad(½).

  By CLT, \( \frac{1}{n} \langle σ, σ̄ \rangle = \frac{1}{n} \sum_{i=1}^{n} σ_i \sigmā_i \xrightarrow{D} \mathcal{N}(0,1) \)

  Asymptotic convergence of MGF (see Lemma 3 below), we have.

  \[ E_\theta, \alpha \left[ e^{\frac{2(0)(\sigma, \sigmā)}{2n}} \right] \to E[I \left[ e^{\frac{2(0)(\alpha, \alpha)}{2}} \right] \]

  \[ = \begin{cases} 
  0 & \text{if } \alpha > 1 \\
  \frac{1}{\alpha(1)} & \text{if } \alpha < 1 
  \end{cases} \]

  MGF of \( \chi^2(1) \) distribution.

- Exact bisection.

  Let us consider the case σ, σ̄ are drawn iid and uniformly at random from the set \( \{ \theta \in \mathbb{R}^n : \sum \theta_i = 0 \} \).

  For simplicity, let

  \( \sigma = 2^{3} - 1 \) and \( \sigmā = 2^{3} - 1 \)

  Then \( \langle σ, σ̄ \rangle = 4 \langle 3, 3 \rangle - n \)

  and \( \langle 3, 3 \rangle \sim \text{Hypergeom} \left( n, \frac{n}{2}, \frac{n}{2} \right) \)

  Thus \( \langle 3, 3 \rangle - \frac{4}{n} \xrightarrow{D} \mathcal{N}(0,1) \).

  Therefore,
\[ E_{\theta, \delta} \left[ e^{\frac{-2\theta(x)}{2\delta^2}} \right] \]
\[ = E_{\theta, \delta} \left[ e^{\frac{-2\theta(x)}{2\delta^2}} \left( 4\langle 3, \frac{x}{2\delta^2} \rangle - n \right)^2 \right] \]
\[ \rightarrow E \left[ e^{\frac{-2\theta(x)}{2} x^2} \right] = \gamma + \infty \text{ if } n \geq 1 \]
\[ O(1), \text{ if } n < 1 \]

**Lemma 3.** [Convergence of MGF]

Assume \( X_n \overset{D}{\to} X \) and \( M_n(t) = E[e^{tX_n}] \) and \( M(t) = E[e^{tX}] \)

If there exists a constant \( \lambda > 0 \) s.t.

\[ \sup_n \left| P \left\{ |X_n| > x \right\} \right| \leq e^{-\lambda x} \text{ for all } x > 0, \]

then \( M_n(t) \to M(t) \) for all \( |t| < \lambda \).

**pf:** Uniform integrability. Exercise

**Remark:** The critical threshold \( \frac{(a-b)^2}{2(\omega^2)} = 1 \) also implies non-detect.

Proof of this is outside the scope.

- The threshold of the spiked Wigner model is given by

\[ \lambda = 1 \text{ using the second moment method (HW).} \]