# Hypothesis Testing Problem in Stochastic Block Model

Feng Zhao

June 7, 2021

# Hypothesis Testing

Stardard hypothesis testing problem:

$$\begin{cases} H_0 : X \sim P_0 \\ H_1 : X \sim P_1 \end{cases}$$

- Discrete alphabet of $X$: $\mathcal{X}$
- $n$ i.i.d. observations $x^{(n)} = (x_1, \ldots, x_n)$
- The averaged error

$$P_e := P(\widehat{H} = 1|H_0)P(H_0) + P(\widehat{H} = 0|H_1)P(H_1)$$

- Chernoff information for optimal test $\widehat{H}$

$$-\lim_{n\to\infty} \frac{1}{n} \log P_e = -\min_{\lambda\in[0,1]} \log \sum_{x\in\mathcal{X}} P_0^{1-\lambda}(x) P_1^{\lambda}(x)$$

# Hypothesis Testing

Paired hypothesis testing problem:

$$\begin{cases} H_0 : X \sim P_0 = (P \times Q) \\ H_1 : X \sim P_1 = (Q \times P) \end{cases}$$

- Random variable $X = (X_1, X_2)$: $X_1 \sim P, X_2 \sim Q, X_1 \perp\!\!\!\perp X_2$
- Chernoff information for optimal test $\widehat{H}$

$$-\lim_{n\to\infty} \frac{1}{n} \log P_e = -\min_{\lambda\in[0,1]} \log \sum_{x,y\in\mathcal{X}} P_0^{1-\lambda}(x,y) P_1^{\lambda}(x,y)$$

$$= -\min_{\lambda\in[0,1]} \left( \log \sum_{x\in\mathcal{X}} P^{1-\lambda}(x) Q^{\lambda}(x) + \log \sum_{y\in\mathcal{X}} Q^{1-\lambda}(y) P^{\lambda}(y) \right)$$

- $\lambda = \frac{1}{2}$: minimizer

Paired hypothesis testing problem:

$$\begin{cases} H_0 : X \sim P_0 = (P \times Q) \\ H_1 : X \sim P_1 = (Q \times P) \end{cases}$$

Chernoff information for optimal test $\widehat{H}$

$$-\lim_{n \to \infty} \frac{1}{n} \log P_e = -2 \log \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)}$$

Rényi divergence with order $\frac{1}{2}$

$$D_{1/2}(P||Q) := -2 \log \sum_{x \in \mathcal{X}} \sqrt{P(x)Q(x)}$$

# Stochastic Block Model

A probabilistic model to generate random graph

- $Y_i$: label for the $i$-th node
- $X_{ij} = 1$: an edge exists between node $i$ and $j$

Procedures:

1. Generate $Y_1, \ldots, Y_n$ uniformly from $\{\pm 1\}^n$
2. Make sure $\sum_{i=1}^{n} Y_i = 0$
3. $X_{ij} \sim \mathrm{Bern}(p)$ if $Y_i = Y_j$
4. $X_{ij} \sim \mathrm{Bern}(q)$ if $Y_i \neq Y_j$

Misclassfication of label of one node

- $Y_3, \ldots, Y_n$ are given, satisfying $Y_3 + \cdots + Y_n = 0$
- What's the error rate of the optimal estimator for $Y_1$ and $Y_2$?

# Hypothesis Testing in Stochastic Block Model

Paired hypothesis testing problem:

$$\begin{cases} H_0 : Y_1 = 1 \text{ and } Y_2 = -1 \iff X \sim \mathrm{Bern}(p) \times \mathrm{Bern}(q) \\ H_1 : Y_1 = -1 \text{ and } Y_2 = 1 \iff X \sim \mathrm{Bern}(q) \times \mathrm{Bern}(p) \end{cases}$$

- $n-1$ i.i.d. observations of $X$
- Chernoff information for optimal test $\widehat{H}$

$$-\lim_{n\to\infty} \frac{1}{n} \log P_e = -2\log(\sqrt{pq} + \sqrt{(1-p)(1-q)})$$

- What if $p, q$ varies with $n$?

## Theorem (Cramér Theorem)

$X_1, \dots, X_n$ *i.i.d.* $\sim P$, $\gamma > \mathbb{E}[X_1]$,

$$-\lim_{n\to\infty} \frac{1}{n} \log P\left(\frac{X_1 + \cdots + X_n}{n} > \gamma\right) = \psi_P^*(\gamma)$$

Chernoff Information

- $X, X_1, \dots, X_n$ i.i.d. $\sim P_0$
- $\ell(X) = \log \frac{P_1(X)}{P_0(X)} \sim P$

$$-\lim_{n\to\infty} \frac{1}{n} \log P_e = \psi_P^*(0)$$

## Theorem (Gärtner Ellis Theorem)

$X, X_1, \ldots, X_n$ *i.i.d.* $\sim P_n$, $\gamma > \lim_{n\to\infty} \frac{n}{\gamma_n}\mathbb{E}[X_1]$,

$$-\lim_{n\to\infty} \frac{1}{\gamma_n} \log P\left(\frac{X_1 + \cdots + X_n}{\gamma_n} > \gamma\right) = \psi_P^*(\gamma)$$

- $\lim_{n\to\infty} \gamma_n = +\infty$
- *Distribution* $P_n$ *depends on* $n$
- *log-MGF:* $\psi_P(\lambda) = \lim_{n\to\infty} \frac{n}{\gamma_n} \log \mathbb{E}[e^{\lambda X}]$

## Chernoff Information

- $X, X_1, \ldots, X_n$ i.i.d. $\sim P_{0,n}$
- $\ell(X) = \log \frac{P_{1,n}(X)}{P_{0,n}(X)} \sim P_n$

$$-\lim_{n\to\infty} \frac{1}{\gamma_n} \log P_e = \psi_P^*(0)$$

# Stochastic Block Model

- $Y_i$: label for the $i$-th node
- $X_{ij} = 1$: an edge exists between node $i$ and $j$
- $p_n = \frac{a \log n}{n}, q_n = \frac{b \log n}{n}$

Procedures:

1. Generate $Y_1, \ldots, Y_n$ uniformly from $\{\pm 1\}^n$
2. Make sure $\sum_{i=1}^n Y_i = 0$
3. $X_{ij} \sim \mathrm{Bern}(p_n)$ if $Y_i = Y_j$
4. $X_{ij} \sim \mathrm{Bern}(q_n)$ if $Y_i \neq Y_j$

Misclassfication of label of one node

- $Y_3, \ldots, Y_n$ are given, satisfying $Y_3 + \cdots + Y_n = 0$
- What's the error rate of the optimal estimator for $Y_1$ and $Y_2$?

# Hypothesis Testing in Stochastic Block Model

Paired hypothesis testing problem:

$$\begin{cases} H_0 : Y_1 = 1 \text{ and } Y_2 = -1 \iff X \sim P_{0,n} = \mathrm{Bern}(p_n) \times \mathrm{Bern}(q_n) \\ H_1 : Y_1 = -1 \text{ and } Y_2 = 1 \iff X \sim P_{1,n} = \mathrm{Bern}(q_n) \times \mathrm{Bern}(p_n) \end{cases}$$

Choose $\gamma_n = \log n$

$$\begin{aligned} \psi_P(\lambda) &= \lim_{n\to\infty} \frac{n}{\log n} \log E_{P_{0,n}}[e^{\lambda \ell(X)}] \\ &= a^{1-\lambda} b^{\lambda} + a^{\lambda} b^{1-\lambda} - a - b \end{aligned}$$

Polynomial error rate

$$-\lim_{n\to\infty} \frac{1}{\log n} \log P_e = -\min_{\lambda} \psi_P(\lambda) = (\sqrt{a} - \sqrt{b})^2$$

# Conclusion

- In paired hypothesis testing, Chernoff information $\Rightarrow$ Rényi divergence with order $\frac{1}{2}$
- Gärtner Ellis Theorem generalizes Cramér Theorem, allowing the derivation of polynomial error rate

# Questions and Answers