Low coordinate degree algorithms II: Categorical signals and generalized stochastic block models

Dmitriy (Tim) Kunisky

Department of Applied Mathematics & Statistics, Johns Hopkins University

COLT 2025: Session on Computational Complexity
July 4, 2025

Hypothesis Testing

Task: Distinguish high-dimensional probability measures
$\mathbb{Q}_N$ (null) vs. $\mathbb{P}_N$ (alternative/planted) on $\mathbb{R}^N$.

Hypothesis Testing

Task: Distinguish high-dimensional probability measures $\mathbb{Q}_N$ (null) vs. $\mathbb{P}_N$ (alternative/planted) on $\mathbb{R}^N$.

Namely, find **tests** $t_N : \mathbb{R}^N \to \{\mathsf{q}, \mathsf{p}\}$ such that

$$\lim_{N\to\infty} \underbrace{\Pr_{\boldsymbol{Y}\sim\mathbb{Q}_N}[t_N(\boldsymbol{Y}) = \mathsf{p}]}_{\text{Type I error}} = 0,$$

$$\lim_{N\to\infty} \underbrace{\Pr_{\boldsymbol{Y}\sim\mathbb{P}_N}[t_N(\boldsymbol{Y}) = \mathsf{q}]}_{\text{Type II error}} = 0.$$

Hypothesis Testing

Task: Distinguish high-dimensional probability measures $\mathbb{Q}_N$ (null) vs. $\mathbb{P}_N$ (alternative/planted) on $\mathbb{R}^N$.

Namely, find **tests** $t_N : \mathbb{R}^N \to \{\mathsf{q}, \mathsf{p}\}$ such that

$$\lim_{N \to \infty} \underbrace{\Pr_{\boldsymbol{Y} \sim \mathbb{Q}_N} [t_N(\boldsymbol{Y}) = \mathsf{p}]}_{\text{Type I error}} = 0,$$

$$\lim_{N \to \infty} \underbrace{\Pr_{\boldsymbol{Y} \sim \mathbb{P}_N} [t_N(\boldsymbol{Y}) = \mathsf{q}]}_{\text{Type II error}} = 0.$$

When do such t_N achieving **(strong) detection** exist?
When are they computable in **polynomial time**?

Equivalent Formulation: Separating Statistics

Definition: $f_N : \mathbb{R}^N \to \mathbb{R}$ achieve **separation** if

$$\mathop{\mathbb{E}}_{\boldsymbol{Y}\sim\mathbb{P}_N} f_N(\boldsymbol{Y}) - \mathop{\mathbb{E}}_{\boldsymbol{Y}\sim\mathbb{Q}_N} f_N(\boldsymbol{Y}) \gg \sqrt{\mathop{\mathsf{Var}}_{\boldsymbol{Y}\sim\mathbb{P}_N} f_N(\boldsymbol{Y})} + \sqrt{\mathop{\mathsf{Var}}_{\boldsymbol{Y}\sim\mathbb{Q}_N} f_N(\boldsymbol{Y})}.$$

Equivalent Formulation: Separating Statistics

Definition: $f_N : \mathbb{R}^N \to \mathbb{R}$ achieve **separation** if

$$\mathop{\mathbb{E}}_{\boldsymbol{Y} \sim \mathbb{P}_N} f_N(\boldsymbol{Y}) - \mathop{\mathbb{E}}_{\boldsymbol{Y} \sim \mathbb{Q}_N} f_N(\boldsymbol{Y}) \gg \sqrt{\mathop{\mathsf{Var}}_{\boldsymbol{Y} \sim \mathbb{P}_N} f_N(\boldsymbol{Y})} + \sqrt{\mathop{\mathsf{Var}}_{\boldsymbol{Y} \sim \mathbb{Q}_N} f_N(\boldsymbol{Y})}.$$

Proposition: Detection $\Leftrightarrow$ separation (and equivalence preserves runtime).

If t_N detect, then $f_N(\boldsymbol{Y}) := \mathbb{1}\{t_N(\boldsymbol{Y}) = \mathsf{p}\}$ separate.

If f_N separate, set $\eta := \frac{1}{2}(\mathbb{E}_{\boldsymbol{Y} \sim \mathbb{P}_N} f_N(\boldsymbol{Y}) + \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}_N} f_N(\boldsymbol{Y}))$. Then, by Chebyshev's inequality, thresholding test will detect:

$$t_N(\boldsymbol{Y}) := \left\{ \begin{array}{ll} \mathsf{p} & \text{if } f_N(\boldsymbol{Y}) \geq \eta \\ \mathsf{q} & \text{if } f_N(\boldsymbol{Y}) < \eta \end{array} \right\}.$$

Example: Stochastic Block Model (SBM)

Specific widely studied example: $N = \binom{n}{2}$, and

1. $\mathbb{Q}_N$: $Y_{ij} \overset{\text{(iid)}}{\sim} \mathsf{Ber}(\frac{a+(k-1)b}{kn})$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k])$,
   $Y_{ij} \overset{\text{(ind)}}{\sim} \mathsf{Ber}(\frac{a}{n})$ if $x_i = x_j$ and $\mathsf{Ber}(\frac{b}{n})$ if $x_i \neq x_j$.

Example: Stochastic Block Model (SBM)

Specific widely studied example: $N = \binom{n}{2}$, and

1. $\mathbb{Q}_N$: $Y_{ij} \overset{\text{(iid)}}{\sim} \mathsf{Ber}(\frac{a+(k-1)b}{kn})$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k])$,

   $Y_{ij} \overset{\text{(ind)}}{\sim} \mathsf{Ber}(\frac{a}{n})$ if $x_i = x_j$ and $\mathsf{Ber}(\frac{b}{n})$ if $x_i \neq x_j$.

Conjecture: [DKMZ '11] If

$$\frac{(a-b)^2}{k(a+(k-1)b)} < 1, \tag{KS}$$

then no polynomial-time algorithm can detect.

Example: Stochastic Block Model (SBM)

Specific widely studied example: $N = \binom{n}{2}$, and

1. $\mathbb{Q}_N$: $Y_{ij} \overset{\text{(iid)}}{\sim} \mathsf{Ber}(\frac{a+(k-1)b}{kn})$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k])$,
   $Y_{ij} \overset{\text{(ind)}}{\sim} \mathsf{Ber}(\frac{a}{n})$ if $x_i = x_j$ and $\mathsf{Ber}(\frac{b}{n})$ if $x_i \neq x_j$.

Conjecture: [DKMZ '11] If

$$\frac{(a-b)^2}{k(a+(k-1)b)} < 1, \tag{KS}$$

then no polynomial-time algorithm can detect.

Evidence: [HS '17], [Hopkins '18], [BBKMW '20] Under KS, no **polynomials** of **low degree** $\leq O(n/\log n)$ can separate.

Generalized Stochastic Block Models

Generic model of independent p-wise interactions of n individuals: for some μ and $\nu_\bullet$ measures on **arbitrary** Ω,

1. $\mathbb{Q}_N$: $Y_{\{i_1,\dots,i_p\}} \overset{\text{(iid)}}{\sim} \mu$.
2. $\mathbb{P}_N$: $x_1, \dots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k]) \rightsquigarrow Y_{\{i_1,\dots,i_p\}} \overset{\text{(ind)}}{\sim} \nu_{\{x_{i_1},\dots,x_{i_p}\}}$.

Generalized Stochastic Block Models

Generic model of independent p-wise interactions of n individuals: for some μ and $\nu_\bullet$ measures on **arbitrary** Ω,

1. $\mathbb{Q}_N : Y_{\{i_1,\ldots,i_p\}} \overset{\text{(iid)}}{\sim} \mu$.
2. $\mathbb{P}_N : x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k]) \rightsquigarrow Y_{\{i_1,\ldots,i_p\}} \overset{\text{(ind)}}{\sim} \nu_{\{x_{i_1},\ldots,x_{i_p}\}}$.

Example: To recreate two-community ($k = 2$) graph SBM,

- $p = 2$ (binary interactions)
- $\Omega = \{0, 1\}$ (binary observations)
- $\mu = \mathsf{Ber}(\frac{a+b}{2n})$, $\nu_{\{1,2\}} = \mathsf{Ber}(\frac{a}{n})$, $\nu_{\{1,1\}} = \nu_{\{2,2\}} = \mathsf{Ber}(\frac{b}{n})$

Generalized Stochastic Block Models

Generic model of independent p-wise interactions of n individuals: for some μ and $\nu_\bullet$ measures on **arbitrary** Ω,

1. $\mathbb{Q}_N$: $Y_{\{i_1,\dots,i_p\}} \overset{\text{(iid)}}{\sim} \mu$.
2. $\mathbb{P}_N$: $x_1, \dots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k]) \rightsquigarrow Y_{\{i_1,\dots,i_p\}} \overset{\text{(ind)}}{\sim} \nu_{\{x_{i_1},\dots,x_{i_p}\}}$.

Example: To recreate two-community ($k = 2$) graph SBM,

- $p = 2$ (binary interactions)
- $\Omega = \{0, 1\}$ (binary observations)
- $\mu = \mathsf{Ber}(\frac{a+b}{2n})$, $\nu_{\{1,2\}} = \mathsf{Ber}(\frac{a}{n})$, $\nu_{\{1,1\}} = \nu_{\{2,2\}} = \mathsf{Ber}(\frac{b}{n})$

Generalizations:

- $p > 2 \rightsquigarrow$ **hypergraph SBM**
- $|\Omega| > 2 \rightsquigarrow$ **labelled SBM** (can be infinite!)

Low Coordinate Degree Functions [Hopkins '18]

For Ω non-numerical, cannot use polynomials, but...

Low Coordinate Degree Functions [Hopkins '18]

For Ω non-numerical, cannot use polynomials, but...
functions of **coordinate degree** $\leq D$:

$$V_{\leq D} := \left\{ \sum_{\substack{T \subseteq [N] \\ |T| \leq D}} f_T(\boldsymbol{Y}) : f_T(\boldsymbol{Y}) \text{ depends only on } (Y_i)_{i \in T} \right\}.$$

$\mathbb{R}_{\leq D}[\boldsymbol{Y}] \subseteq V_{\leq D} \rightsquigarrow$ larger class of candidate statistics. E.g., any **coordinate decision tree** of depth $\leq D$.

Low Coordinate Degree Functions [Hopkins ’18]

For Ω non-numerical, cannot use polynomials, but...
functions of **coordinate degree** $\leq D$:

$$V_{\leq D} := \left\{ \sum_{\substack{T \subseteq [N] \\ |T| \leq D}} f_T(\boldsymbol{Y}) : f_T(\boldsymbol{Y}) \text{ depends only on } (Y_i)_{i \in T} \right\}.$$

$\mathbb{R}_{\leq D}[\boldsymbol{Y}] \subseteq V_{\leq D} \rightsquigarrow$ larger class of candidate statistics. E.g., any **coordinate decision tree** of depth $\leq D$.

Fact: On $\{0, 1\}^N$, degree and coordinate degree coincide.

Low Coordinate Degree Functions [Hopkins '18]

For Ω non-numerical, cannot use polynomials, but...
functions of **coordinate degree** $\le D$:

$$V_{\le D} := \left\{ \sum_{\substack{T \subseteq [N] \\ |T| \le D}} f_T(\boldsymbol{Y}) : f_T(\boldsymbol{Y}) \text{ depends only on } (Y_i)_{i \in T} \right\}.$$

$\mathbb{R}_{\le D}[\boldsymbol{Y}] \subseteq V_{\le D} \rightsquigarrow$ larger class of candidate statistics. E.g., any **coordinate decision tree** of depth $\le D$.

Fact: On $\{0, 1\}^N$, degree and coordinate degree coincide.

Natural generalization of polynomials to product domains. Main proof technique of **Boolean Fourier analysis** for low-degree polynomials also conveniently generalizes to **Efron-Stein/Hoeffding/ANOVA decomposition**.

General-Purpose Lower Bound Condition

Theorem: [K. '24] Let $D = D(n) \in \mathbb{N}$. In any generalized stochastic block model (subject to mild symmetry constraint), if

$$\mathop{\mathbb{E}}_{\boldsymbol{r} \sim \mathrm{Mult}(n,k^2)} \sum_{d=0}^{D(n)} \frac{1}{d!} \langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle^d = O(1),$$

then coordinate degree $\leq D(n)$ functions cannot separate.

General-Purpose Lower Bound Condition

Theorem: [K. '24] Let $D = D(n) \in \mathbb{N}$. In any generalized stochastic block model (subject to mild symmetry constraint), if

$$\mathop{\mathbb{E}}_{\boldsymbol{r} \sim \mathrm{Mult}(n,k^2)} \sum_{d=0}^{D(n)} \frac{1}{d!} \langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle^d = O(1),$$

then coordinate degree $\leq D(n)$ functions cannot separate.

Here $\boldsymbol{T} \in (\mathbb{R}^{[k]\times[k]})^{\otimes p}$ is a **fixed** tensor constructed from the model parameters and measuring "effective noisiness":

$$T_{(a_1,b_1),\dots(a_p,b_p)} = \frac{1}{p!} \cdot \left(\mathop{\mathbb{E}}_{y \sim \mu} \frac{d\nu_{\{a_1,\dots,a_p\}}}{d\mu}(y) \frac{d\nu_{\{b_1,\dots,b_p\}}}{d\mu}(y) - 1 \right).$$

General-Purpose Lower Bound Condition

Theorem: [K. '24] Let $D = D(n) \in \mathbb{N}$. In any generalized stochastic block model (subject to mild symmetry constraint), if

$$\mathop{\mathbb{E}}_{\boldsymbol{r} \sim \mathsf{Mult}(n,k^2)} \sum_{d=0}^{D(n)} \frac{1}{d!} \langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle^d = O(1),$$

then coordinate degree $\leq D(n)$ functions cannot separate.

Here $\boldsymbol{T} \in (\mathbb{R}^{[k]\times[k]})^{\otimes p}$ is a **fixed** tensor constructed from the model parameters and measuring "effective noisiness":

$$T_{(a_1,b_1),\dots(a_p,b_p)} = \frac{1}{p!} \cdot \left(\mathop{\mathbb{E}}_{y\sim\mu} \frac{d\nu_{\{a_1,\dots,a_p\}}}{d\mu}(y) \frac{d\nu_{\{b_1,\dots,b_p\}}}{d\mu}(y) - 1 \right).$$

Instructive special case:

$$T_{(a_1,a_1),\dots,(a_p,a_p)} = \frac{1}{p!} \cdot \chi^2(\nu_{a_1,\dots,a_p} \,\|\, \mu).$$

Using the General-Purpose Lower Bound

Main point: Reduces proving **high-dimensional** lower bounds to studying random **scalar**

$$\langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle,$$

and thus in turn to (the moments of) the multinomial random vector

$$\boldsymbol{r} \sim \mathsf{Mult}(n, k^2).$$

Using the General-Purpose Lower Bound

Main point: Reduces proving **high-dimensional** lower bounds to studying random **scalar**

$$\langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle,$$

and thus in turn to (the moments of) the multinomial random vector

$$\boldsymbol{r} \sim \mathsf{Mult}(n, k^2).$$

For applications, a slightly more careful version of a very naive bound suffices:

$$|\langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle| \leq \|\boldsymbol{T}\|_{\mathsf{inj}} \|\boldsymbol{r}\|_2^p.$$

Using the General-Purpose Lower Bound

Main point: Reduces proving **high-dimensional** lower bounds to studying random **scalar**

$$\langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle,$$

and thus in turn to (the moments of) the multinomial random vector

$$\boldsymbol{r} \sim \mathsf{Mult}(n, k^2).$$

For applications, a slightly more careful version of a very naive bound suffices:

$$|\langle \boldsymbol{T}, \boldsymbol{r}^{\otimes p} \rangle| \leq \|\boldsymbol{T}\|_{\mathsf{inj}} \|\boldsymbol{r}\|_2^p.$$

Actually, need to first decompose

$$\boldsymbol{r} = \mathbb{E}\boldsymbol{r} + (\boldsymbol{r} - \mathbb{E}\boldsymbol{r}) = \frac{n}{k^2}\mathbf{1} + \overline{\boldsymbol{r}}.$$

Application 1: Hypergraph SBMs

For $p > 2$,

1. $\mathbb{Q}_N$: $Y_{\{i_1,\ldots,i_p\}} \overset{\text{(iid)}}{\sim} \mathsf{Ber}(\frac{a+(k^{p-1}-1)b}{k^{p-1}n})$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k])$,

   $Y_{\{i_1,\ldots,i_p\}} \overset{\text{(ind)}}{\sim} \mathsf{Ber}(\frac{a}{n})$ if $x_{i_1} = \cdots = x_{i_p}$, $\mathsf{Ber}(\frac{b}{n})$ else.

Application 1: Hypergraph SBMs

For $p > 2$,

1. $\mathbb{Q}_N$: $Y_{\{i_1,\dots,i_p\}} \overset{\text{(iid)}}{\sim} \mathsf{Ber}\left(\frac{a+(k^{p-1}-1)b}{k^{p-1}n}\right)$.
2. $\mathbb{P}_N$: $x_1, \dots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k])$,
   $Y_{\{i_1,\dots,i_p\}} \overset{\text{(ind)}}{\sim} \mathsf{Ber}\left(\frac{a}{n}\right)$ if $x_{i_1} = \cdots = x_{i_p}$, $\mathsf{Ber}\left(\frac{b}{n}\right)$ else.

Conjecture: [ACKZ '15] If

$$\frac{(p-1)(a-b)^2}{k^{p-1}(a+(k^{p-1}-1)b)} < 1, \tag{KS}$$

then no polynomial-time algorithm can detect.

Application 1: Hypergraph SBMs

For $p > 2$,

1. $\mathbb{Q}_N$: $Y_{\{i_1,\ldots,i_p\}} \overset{\text{(iid)}}{\sim} \mathsf{Ber}(\frac{a+(k^{p-1}-1)b}{k^{p-1}n})$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}([k])$,
   $Y_{\{i_1,\ldots,i_p\}} \overset{\text{(ind)}}{\sim} \mathsf{Ber}(\frac{a}{n})$ if $x_{i_1} = \cdots = x_{i_p}$, $\mathsf{Ber}(\frac{b}{n})$ else.

Conjecture: [ACKZ '15] If

$$\frac{(p-1)(a-b)^2}{k^{p-1}(a+(k^{p-1}-1)b)} < 1, \tag{KS}$$

then no polynomial-time algorithm can detect.

Theorem: [K. '24] Under KS, no functions (polynomials) of coordinate degree (degree) $\leq O(n/\log n)$ can separate.

Application 2: Group Synchronization

For $p = 2$ and G a finite group,

1. $\mathbb{Q}_N$: $Y_{ij} \overset{\text{(iid)}}{\sim} \mathsf{Unif}(G)$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}(G)$,

$$Y_{ij} \overset{\text{(ind)}}{\sim} \left\{ \begin{array}{ll} x_i x_j^{-1} & \text{with probability } \frac{q}{\sqrt{n}} \\ \mathsf{Unif}(G) & \text{with probability } 1 - \frac{q}{\sqrt{n}} \end{array} \right\}.$$

Application 2: Group Synchronization

For $p = 2$ and G a finite group,

1. $\mathbb{Q}_N$: $Y_{ij} \overset{\text{(iid)}}{\sim} \mathsf{Unif}(G)$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}(G)$,

$$Y_{ij} \overset{\text{(ind)}}{\sim} \left\{ \begin{array}{ll} x_i x_j^{-1} & \text{with probability } \frac{q}{\sqrt{n}} \\ \mathsf{Unif}(G) & \text{with probability } 1 - \frac{q}{\sqrt{n}} \end{array} \right\}.$$

Conjecture: [Singer '11, PWBM '16] If $q < 1$, then no polynomial-time algorithm can detect.

(For $G = \mathbb{Z}/k\mathbb{Z}$, matches performance of simple spectral algorithm.)

Application 2: Group Synchronization

For $p = 2$ and G a finite group,

1. $\mathbb{Q}_N$: $Y_{ij} \overset{\text{(iid)}}{\sim} \mathsf{Unif}(G)$.
2. $\mathbb{P}_N$: $x_1, \ldots, x_n \overset{\text{(iid)}}{\sim} \mathsf{Unif}(G)$,

$$Y_{ij} \overset{\text{(ind)}}{\sim} \left\{ \begin{array}{ll} x_i x_j^{-1} & \text{with probability } \frac{q}{\sqrt{n}} \\ \mathsf{Unif}(G) & \text{with probability } 1 - \frac{q}{\sqrt{n}} \end{array} \right\}.$$

Conjecture: [Singer '11, PWBM '16] If $q < 1$, then no polynomial-time algorithm can detect.

(For $G = \mathbb{Z}/k\mathbb{Z}$, matches performance of simple spectral algorithm.)

Theorem: [K. '24] If $q < 1$, then no functions of coordinate degree $\leq O(n/\log n)$ can separate.

(Formulation requires coordinate degree!)

Thank you!

Omitted Highlights

- Arbitrary interaction probabilities in graph and hypergraph SBMs; lower bounds match associated generalized KS thresholds.

Omitted Highlights

- Arbitrary interaction probabilities in graph and hypergraph SBMs; lower bounds match associated generalized KS thresholds.
- Lower bounds against subexponential-time algorithms (e.g., for planted XOR-SAT written as generalized SBM).

Omitted Highlights

- Arbitrary interaction probabilities in graph and hypergraph SBMs; lower bounds match associated generalized KS thresholds.

- Lower bounds against subexponential-time algorithms (e.g., for planted XOR-SAT written as generalized SBM).

- Analogy between subexponential-time algorithms for SBMs and for continuous models: each SBM looks, in terms of coarse computational characteristics, like a spiked p_*-tensor model for some p_*.