Learning on graphs : phase transitions.

Thresholds in random graphs

1. **Thresholds in random graphs: connectivity**
   Let $G \sim \mathbb{G}(n,p)$ with $p = \frac{c \log n}{n}$. Show that $\log n/n$ is a sharp threshold for containing an isolated vertex, i.e.,

   (a) if $c < 1$, then G has an isolated vertex whp,

   (b) if $c > 1$, then G does not have an isolated vertex whp.

   Use 1a to show that $\log n/n$ is a sharp threshold for G being connected.

2. Let Z_n denote the number of triangles in $G \sim G(n, \frac{c}{n})$. Show that $Z_n \sim \text{Poi}\left(\frac{c^3}{6}\right)$.

3. **Low-degree EASY region.** Let $\varepsilon > 0$ and $k \geqslant n^{1/2+\varepsilon}$, and consider $H_0 : G \sim G(n, 1/2)$ vs $H_1 : G \sim G(n, k, 1/2)$. Show that the 'signed triangle count'

$$f(G) := \sum_{i<j<k} (A_{ij} - \tfrac{1}{2})(A_{ik} - \tfrac{1}{2})(A_{jk} - \tfrac{1}{2})$$

   strongly separates H_0 vs H_1.

Properties of the planted clique model.

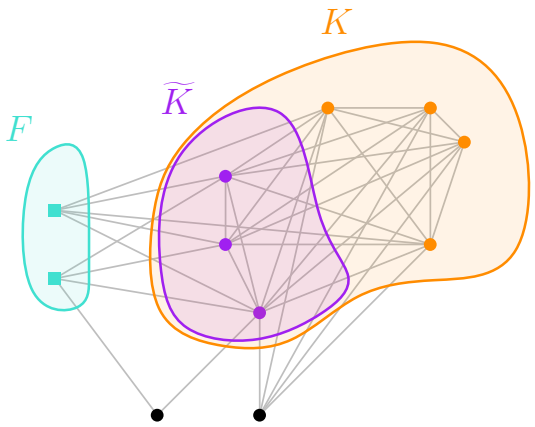


Figure 1: Given any $\tilde{K} \subset K$, where K is the planted clique. Let F be the set of common neighbours of $\tilde{K}$ which are not in the clique K. The idea is to show that for $G \sim G(n, k, 1/2)$ whp the graph has the property that for any subset $\tilde{K} \subset K$ of a certain size, the k maximum degree vertices in $G' = G[K \cup F]$ are the vertices of the planted clique K. See Question 7.

5. **PC and low-degree framework** Let $H_1 : G \sim G(n, k, 1/2)$, and $h_\alpha(G) = \prod_{ij \in \alpha}(2A_{ij} - 1)$. Prove that $\mathbb{E}_1[h_\alpha(G)] = \left(\frac{k}{n}\right)^{|V(\alpha)|}$.

6. **Maximum degrees to find planted clique** Suppose that $G \sim G'(n, k, 1/2)$ with planted clique K. Let $\hat{K} := \{k \text{ vertices of highest degree in } G\}$.

   (a) For vertex i, give the distribution of the degree d_i, conditioned on $i \notin K$.

   (b) (As above), for $i \in K$.

(c) Prove that

$$\mathbb{P}[K = \hat{K}] \geqslant 1 - 2n \exp\left(\frac{-k^2}{2n}\right)$$

7. **Towards fast algorithms for finding planted clique above the threshold.**

   For a graph $g = ([n], E(g))$ and vertex subset $S \subset [n]$ define the set of common neighbours to be $C_g(S) = \{i \in [n] \,:\, ij \in E(g) \text{ for all } j \in S\}$. Prove the following.

   Fix $\varepsilon > 0$, $s = \lceil (1+\varepsilon) \log_2 n \rceil$, $k \geqslant C \log n$ for some $C \geqslant \log n$. Let $G \sim G'(n, k, 1/2)$ with planted set K. Then whp G is such that, for any $\tilde{K} \subset K$ with $|\tilde{K}| = s$ the following holds.

   We have $\hat{K} = K$ where $\hat{K} := \{k \text{ vertices of highest degree in } G[K']\}$ where $K' = \tilde{K} \cup C_G(\tilde{K})$.

   Possible steps.

   (i) (i) Define $F = F(\tilde{K}) = C_G(\tilde{K}) \setminus K$. Show that for fixed $\tilde{K}$ we have ? .

   (ii) By (i) and union bound, given a whp upper bound $\ell = C \log n$ on the size of F. I.e. show that whp G is such that for any $\tilde{K} \subset K$ of size $|\tilde{K}| = s$, then $|F(\tilde{K})| \leqslant \ell$.

   (iii) Show that[1] for $v \in F$, the degree $d_v \sim |\tilde{K}| + \text{Binom}(|K \setminus \tilde{K}|, 1/2)$.

   (iv) Deduce that for $v \in F$, likely that $d_v < k - 1$.

Guassian planted submatrix model

Define the Gaussian planted submatrix model. Sample $Y \sim \text{BC}(n, k, \lambda)$ with planted set K as follows. For each $i \in [n] := \{1, 2, \ldots, n\}$ independently, $i \in K$ with probability k/n. Write $X_{ij} = \mathbf{1}[i, j \in K]$. Independently of X, and for each i, j independently let $Z_{ij} \sim N(0, 1)$. Then $Y = \lambda X + Z$. Define $Y \sim \text{BC}'(n, k, \lambda)$ as above except we take K uniformly over $\binom{[n]}{k}$.

8. Fix α, β such that $0 < \alpha, \beta < 1$ and $\beta > \alpha/2 + 1/2$ (i.e. the green region in Fig 2). Consider $H_0 : Y \sim \text{BC}'(n, 0, 0)$ vs $H_1 : Y \sim \text{BC}'(n, k = n^\beta, \lambda = n^{-\alpha})$.

   Show that the strong detection problem is EASY for α, β. (Find a test ϕ_n for which $r(\phi_n) \to 0$, such that ϕ_n is fast to compute.)

9. Fix α, β such that $0 < \alpha, \beta < 1$ and $\beta > \alpha/2+1/2$ or $\beta > 2\alpha$ (i.e. the purple dashed region and the blue region in Fig 2(b)). Consider $H_0 : Y \sim \text{BC}'(n, 0, 0)$ vs $H_1 : Y \sim \text{BC}'(n, k = n^\beta, \lambda = n^{-\alpha})$.

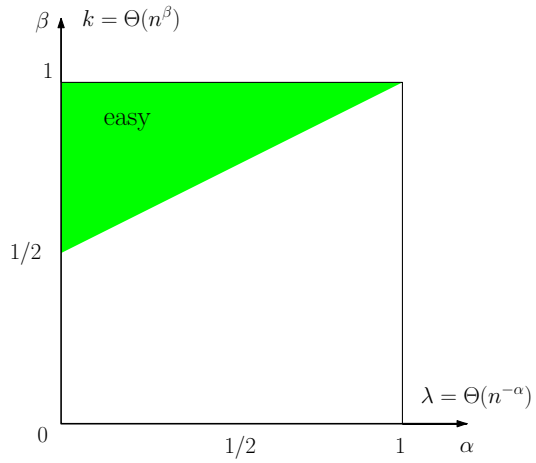
   Show that the strong detection problem is POSSIBLE for α, β. (Find a test ϕ_n for which $r(\phi_n) \to 0$.)

10. Fix α, β such that $\alpha > 0$ and $\beta < \alpha/2 + 1/2$. Consider $H_0 : Y \sim \text{BC}'(n, 0, 0)$ vs $H_1 : Y \sim \text{BC}'(n, k = n^\beta, \lambda = n^{-\alpha})$. Show that no polynomial of degree $O(\log n)$ strongly separates H_0 vs H_1.

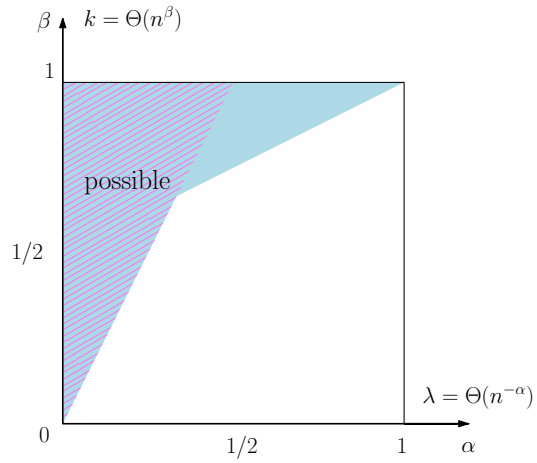
    Hint. There exists a basis[2] $\{h_\alpha\}_\alpha$ for which $\mathbb{E}_0[h_\alpha h_\beta] = \mathbf{1}_{\alpha=\beta}$ and $\mathbb{E}_1[h_\alpha(Y)] = \frac{1}{\sqrt{\alpha!}} \mathbb{E}_1[X^\alpha]$.

[1] or similar, typos expected

[2] Hermite polynomials



(a) detection easy



(b) detection possible

Figure 2: Regions considered in Questions 8 and 9.

Theory of low-degree method.

11. **Connection between likelihood ratio and Adv.** Fix a hypothesis testing problem $H_0 : G \sim Q$ vs $H_1 : G \sim P$. Suppose $\{h_\alpha\}_{\mathcal{I}_D}$ is a basis for polynomials in G of degree at most D such that

$$\langle h_\alpha, h_\beta \rangle_0 := \mathbb{E}_0[h_\alpha(G) h_\beta(G)] = \mathbf{1}_{\alpha=\beta}$$

Define 'projection of likelihood ratio'

$$\widetilde{L}(g) := \sum_{\alpha \in \mathcal{I}_D} \langle L, h_\alpha \rangle_0 \, L(g).$$

Prove that $\mathbb{E}_0[\tilde{L}^2] = \mathsf{Adv}_{\leqslant D}$.

12. **No orthonormal basis for H_0.** Let $\mathbb{P}$ and $\mathbb{Q}$ be planted distributions on $\mathbb{R}^N$, i.e., the *null* distribution is no longer pure noise, but contains a planted structure. Define

$$c_\alpha := \mathbb{E}_{\mathbb{P}}[\phi_\alpha(Y)], \qquad c \in \mathbb{R}^{\mathcal{I}},$$
$$M_{\beta\alpha} := \mathbb{E}_{\mathbb{Q}}[\phi_\alpha(Y)\psi_\beta(W)], \qquad M \in \mathbb{R}^{\mathcal{J}\times\mathcal{I}}.$$

Assume the following

- $\{\phi_\alpha\}_{\alpha\in\mathcal{I}}$ is a basis for f.
- $\{\psi_\beta(W)\}_{\beta\in\mathcal{J}}$ is an orthonormal set in $L^2(\mathbb{Q})$, i.e., $\langle \psi_\beta, \psi_{\beta'} \rangle_{L^2(\mathbb{Q})} := \mathbb{E}_{\mathbb{Q}}[\psi_\beta \psi_{\beta'}] = \mathbb{1}\{\beta = \beta'\}$.
- For c and M as defined above, there exists $u = (u_\beta)_{\beta\in\mathcal{J}}$ satisfying

$$\sum_{\beta\in\mathcal{J}} u_\beta M_{\beta\alpha} = c_\alpha, \quad \forall \alpha \in \mathcal{I}.$$

Then,

$$\mathsf{Adv}^2_{\leqslant D}(\mathbb{P}, \mathbb{Q}) \leqslant \sum_{\beta\in\mathcal{J}} u_\beta^2.$$

Hint: (Bessel's inequality) If H is a Hilbert space and (e_k) is an orthonormal collection in H, then $\sum_{k=1}^{\infty} |\langle x, e_k \rangle|^2 \leqslant \|x\|^2$ for all $x \in H$.

Misc

13. **Reductions**. Write $\mathcal{L}(X)$, by to denote the law of X. For P and P' be probability distributions on the same space. Suppose that $X \sim P$, $\mathcal{A}$ is an map/algorithm, we write

$$P \xrightarrow{\mathcal{A}}_{\varepsilon} P' \qquad \text{if} \qquad d_{\mathrm{TV}}(\mathcal{L}\big(\mathcal{A}(X)\big), P') \leqslant \varepsilon$$

Prove the following, which we will use on Thursday. If P, P_1 and P_2 be three probability spaces, and $\mathcal{A}_1$ and $\mathcal{A}_2$ algorithms such that

$$P \xrightarrow{\mathcal{A}_1}_{\varepsilon_1} P_1 \quad \text{and} \quad P_1 \xrightarrow{\mathcal{A}_2}_{\varepsilon_2} P_2.$$

Then

$$P \xrightarrow[\varepsilon_1+\varepsilon_2]{\mathcal{A}_2 \circ \mathcal{A}_1} P_2.$$

A Hermite expansion

Use the standard shift identity for normalized Hermite polynomials. Namely, for $Z \sim N(0,1)$ and deterministic x,

$$h_m(x+Z) = \sum_{r=0}^{m} \sqrt{\frac{r!}{m!}} \binom{m}{r} x^{m-r} h_r(Z).$$

Applying this identity coordinatewise gives

$$H_\alpha(X+Z) = \sum_{0 \leqslant \beta \leqslant \alpha} \sqrt{\frac{\beta!}{\alpha!}} \binom{\alpha}{\beta} X^{\alpha-\beta} H_\beta(Z).$$

Specifically,

$$\begin{aligned} H_\alpha(Y) = \prod_{i \leqslant j} h_{\alpha_{ij}}(X_{ij} + Z_{ij}) &= \prod_{i \leqslant j} \sum_{k=0}^{\alpha_{ij}} \sqrt{\frac{k!}{\alpha_{ij}!}} \binom{\alpha_{ij}}{k} X_{ij}^{\alpha_{ij}-k} h_k(Z_{ij}) \\ &= \sum_{0 \leqslant \beta \leqslant \alpha} \sqrt{\frac{\beta!}{\alpha!}} \binom{\alpha}{\beta} X^{\alpha-\beta} H_\beta(Z). \end{aligned}$$