Problem Set 2

Spectral clustering and community detection

1. Let B be a $n\times n$ real matrix and $\varepsilon \in (0,1/2)$. Then, for any ε-net $\mathcal{N}$ of $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, we have

$$\|B\| := \sup_{x,y\in S^{n-1}} \langle x, By\rangle \leq \frac{1}{1-2\varepsilon} \cdot \max_{x,y\in\mathcal{N}} \langle x, By\rangle.$$

2. (a) Let X be a mean zero bounded random variable and let Y be an independent copy of X. For $\theta \in \mathbb{R}$, show that

$$\mathbb{E}[e^{\theta X}] \leq \mathbb{E}[e^{\theta(X-Y)}].$$

   (b) Let η be an independent symmetric Bernoulli, i.e, $\mathbb{P}(\eta = +1) = \mathbb{P}(\eta = -1) = 1/2$. Use
   (i) $X - Y \overset{d}{=} \eta(X - Y)$
   (ii) $\frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}, x \in \mathbb{R}$ to show that

$$\mathbb{E}[e^{\theta(X-Y)}] \leq \mathbb{E}[e^{\theta^2(X-Y)^2/2}].$$

   (c) Let B be an $n\times n$ (non-symmetric) matrix such that the entries are independent, mean zero, and $|B_{ij}| \leq 1$ for all i, j. Prove that for any $x, y \in \mathbb{R}^n$ with $\|x\|_2 = \|y\|_2 = 1$ and for any $\theta \in \mathbb{R}$,

$$\mathbb{E}[e^{\theta\langle x,By\rangle}] \leq e^{2\theta^2}.$$

   (d) Use part (c) to conclude that for any $u > 0$,

$$\mathbb{P}\big(\langle x, By\rangle \geq u\big) \leq e^{-u^2/8}.$$

3. (a) Let W_k be the number of walks on $\mathbb{Z} = \{0, \pm1, \pm2, \ldots\}$ of length $2k$ starting and ending at 0, i.e.,

$$W_k = \#\big\{(S_0, S_1, \ldots, S_{2k}) : S_0 = S_{2k} = 0, |S_{i+1} - S_i| = 1 \text{ for each } i\big\}.$$

   Argue that $W_k = \binom{2k}{k}$.

   (b) Let B_k be the number of walks on $\mathbb{Z}$ of length $2k$ which starts at 0 and also ends at 0 and touches (or cross) -1 in between. Argue that $B_k = \binom{2k}{k-1}$.
   [Hint. Let $(0, S_0), (1, S_1), \ldots, (2k, S_{2k})$ be the graph of such a walk in $\mathbb{Z}^2$. Flip the portion of the graph across the line $y = -1$ from the time it first hits -1.]

   (c) Let C_k be the number of walks on $\mathbb{Z}$ of length $2k$ which starts and ends at 0 and always stay at or above 0 in between. Show that

$$C_k = \frac{1}{k+1}\binom{2k}{k}.$$

   (d) Let $d \geq 2$ be an integer. Consider the infinite d-ary tree. Show that the total number of walks of length $2k$ on the vertices of this tree which starts and ends at the root is

$$R_k(d) = \frac{1}{k+1}\binom{2k}{k}d^k.$$

   (e) Argue that $\lim_{k\to\infty} R_k(d)^{1/2k} = 2\sqrt{d}$.

(f)* Suppose now that in the infinite tree, the root has D children ($D > d$) and the rest of vertices have d children as before. Let $\widetilde{R}_k(d, D)$ be the total number of walks of length $2k$ in the modified tree starting and ending at the root. Show that

$$D^k \leq \widetilde{R}_k(d, D) \leq D^k(1 + 4d/D)^k.$$

Therefore,

$$\sqrt{D} \leq \liminf_{k\to\infty} \widetilde{R}_k(d, D)^{1/2k} \leq \limsup_{k\to\infty} \widetilde{R}_k(d, D)^{1/2k} \leq \sqrt{D}(1 + 4d/D)^{1/2}.$$

This implies that if D is very large compared to d, then $\widetilde{R}_k(d, D)$ grows like D^k, which shows that a single large degree vertex can have enormous effect on the total number of walks.

4. Let B be the non-backtracking matrix of graph G with n vertices and m edges.

(a) Show that for any two edges (x, y) and (u, v) of G,

$$(B^t)_{x\to y, u\to v} = B_{y\to x, v\to u}.$$

Let P be the $2m \times 2m$ matrix of size with its rows and columns indexed by directed edges such that

$$P_{x\to y, u\to v} = 1_{\{u=y, x=v\}}.$$

Show that $P^t = P$ and $P^2 = I$. Argue that $B^t = PBP$ and consequently, $(B^k)^t = PB^kP$ for any integer $k \geq 1$.

(b) Assume that the minimum degree of G is at least 2. Show that B has n singular values $\deg(v) - 1, v \in V$ and its other $2m - n$ singular values are 1.
[Hint: BB^t is a block-diagonal matrix.]

5. (Kesten-Stigum bound) Consider the two-type Galton-Watson tree as discussed in the lecture.

- The root o is colored red or blue with probability 1/2.
- Recursively, each vertex gives birth to $\mathrm{Poi}(a/2)$ many vertices of the same color and a $\mathrm{Poi}(b/2)$ many vertices of the opposite color. Assume that $a > b > 0$.
- For each vertex v, assign label $\sigma_v = +1$ or -1 depending on whether the color of v is red or blue respectively.

(a) Argue that the distribution of the above labeled tree is same as follows:
Set $d = (a + b)/2$ and $\varepsilon = b/(a + b)$. Consider a Galton-Watson tree with $\mathrm{Poi}(d)$ offspring distribution. Let σ_o, the label of the root, be ± 1 with equal probability 1/2. Given the label of a parent, each of its children gets the same label of the parent with probability $1 - \varepsilon$ and opposite label with probability ε, independently of others.

[Hint: the following fact (known as Poisson splitting) may be useful. Suppose we have N balls where $N \sim \mathrm{Poi}(\lambda)$. Color each ball independently red or blue with probability p and $1-p$. Let N_1 and N_2 be the number of red and blue balls respectively ($N = N_1 + N_2$). Then $N_1 \sim \mathrm{Poi}(\lambda p)$ and $N_2 \sim \mathrm{Poi}(\lambda(1 - p))$. Moreover, N_1 and N_2 are independent.]

(b) Let F_n be the set of vertices at depth n of the tree from the root. Set $\theta = 1 - 2\varepsilon$. Define

$$S_n := \frac{1}{(d\theta)^n} \sum_{v\in F_n} \sigma_n.$$

Show that $\mathbb{E}[S_n|\sigma_0] = \sigma_0$ for all $n \geq 1$.

(c) Show that

$$\mathrm{var}(S_n) \to \begin{cases} \frac{1}{1-(d\theta^2)^{-1}} & \text{if } d\theta^2 > 1 \\ +\infty & \text{if } d\theta^2 \leq 1. \end{cases}$$

(d) Conclude that there exists a positive constant $c = c(a, b) > 0$ such that

$$\mathrm{corr}(S_n, \sigma_0) \to \begin{cases} c & \text{if } d\theta^2 > 1 \\ 0 & \text{if } d\theta^2 \leq 1. \end{cases}$$

Also check that $d\theta^2 > 1$ if and only if $\frac{a-b}{2} > \sqrt{d}$.

In words, above the KS threshold, i.e., when $\frac{a-b}{2} > \sqrt{d}$, the majority vote of the the labels of the leaves is asymptotically positively correlated with the label of the root σ_o. In contrast, the correlation goes to zero below the KS threshold. In fact, in this case it can be shown that no estimator $\hat{\sigma} = \hat{\sigma}(\sigma_x, x \in F_n)$ is asymptotically positively correlated with σ_o.

6. Let $G \sim G(n, p, q)$ with $p > q$ where the labels σ_i are i.i.d. with $\mathbb{P}(\sigma_i = \pm 1) = 1/2$. Let $k \geq 3$ be an integer. Fix k distinct vertices $v_1, v_2, \ldots, v_k$. Show that

$$\mathbb{P}(v_1 \sim v_2 \sim \cdots \sim v_k \sim v_1) = \left(\frac{p+q}{2}\right)^k + \left(\frac{p-q}{2}\right)^k.$$