Optimal cuts of random geometric graphs

Mathew Penrose
(University of Bath, UK)

Analytic and Geometric approaches to Machine Learning
ICMS Workshop, Bath
July 2021

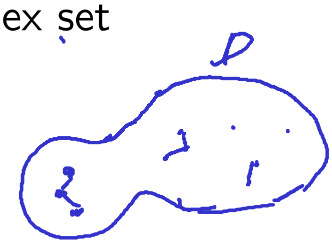
Random geometric graphs (Penrose 2003)

Let D be a bounded region in $\mathbb{R}^d$ (or more generally, a d-dimensional Riemannian manifold) with $d \geq 2$.
Let $X_1, X_2, \ldots, X_n$ be points sampled randomly uniformly from D.

Aim: Learn about D from the sample, via the following graph.

Given $r > 0$, let $G(n, r)$ be the weighted graph on vertex set $V_n := \{X_1, \ldots X_n\}$ with weights

$$W_{xy} := \phi\left(\frac{|x-y|}{r}\right)$$



where $\phi(t) = \mathbf{1}_{[0,1]}(t)$, $t \geq 0$, and $|\cdot|$ is Euclidean.

i.e., connect any two points of V_n at Euclidean distance at most r_n.

[Could also consider non-uniform samples,
and other weight functions ϕ such as $\phi(t) = \exp(-t^2)$]

Isolated vertices of $G(n, r)$

Assume $D \subset \mathbb{R}^d$ open and connected. Also assume D has unit volume and a *Lipschitz boundary* ∂D [this holds e.g. if ∂D is smooth or D is a cube].

Assume we have access to a large sample and can choose $r = r_n, r_n \to 0$. Then asymptotic (large-n) properties of $G(n, r_n)$ may be relevant.

Let $I(G)$ denotes the number of isolated vertices of G,

$\mathbb{E}[I(G(n, r_n)] \sim n \exp(-n\omega_d r_n^d)$ as $n \to \infty$. [$\omega_d :=$ volume of unit ball]

So if $\underbrace{\omega_d n r_n^d}_{\text{av. degree}} = a \log n$ then

$$\lim_{n\to\infty} \mathbb{E}[I(G(n, r_n))] = \begin{cases} \infty \text{ if } a < 1 \\ 0 \text{ if } a > 1. \end{cases}$$

Connectivity of $G(n, r)$

From above: if $\omega_d n r_n^d = a \log n$ [i.e., $r_n = (a \log n/(n\omega_d))^{1/d}$] then $\lim_{n\to\infty} \mathbb{E}[I(G(n, r_n))] = \begin{cases} \infty \text{ if } a < 1 \\ 0 \text{ if } a > 1. \end{cases}$

In fact it turns out that

$$\lim_{n\to\infty} \mathbb{P}[G(n, r_n) \text{ is connected}] = \begin{cases} 0 \text{ if } a < 1 \\ 1 \text{ if } a > 1 \end{cases}$$

If $\omega_d n r_n^d \geq (1 + \varepsilon) \log n$, $G(n, r_n)$ is likely to be connected for large n.

Conversely, if D is *not* connected and $r_n \to 0$, then $G(n, r_n)$ will *not* be connected for large n.
i.e. we can learn about connectivity of D from that of $G(n, r_n)$.

[In preparation with Xiaochuan Yang, exact limit of $\mathbb{P}[G(n, r_n) \text{ connected}]$.

Optimal cuts of a (weighted) graph $G = (V, W)$

For $U \subset V$, set $\partial_G(U) := \sum_{v \in U} \sum_{w \in V \setminus U} W_{vw}$ and $\mathrm{vol}_G(U) := \frac{\#(U)}{\#(V)}$ The **minimum bisection cost** and **cheeger constant (conductance)** of G are

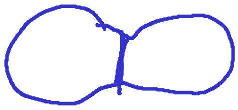
$$\mathrm{MBIS}(G) := \min_{U \subset V : |U| = \lfloor |V|/2 \rfloor} \partial_G(A)$$

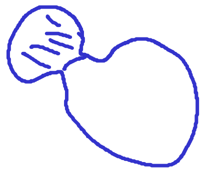
$$\mathrm{CHE}(G) = \min \left\{ \frac{\partial_G(U)}{\mathrm{vol}_G(U)} : U \subset V, 0 < \mathrm{vol}_G(U) \leq 1/2 \right\}$$

The denominator penalizes unbalanced cuts. [Alternatively could define vol(U) by counting edges rather than vertices]
Uses: bounds on mixing times of random walk on graph, bounds on graph laplacian; reasonable criteria for optimal cut.

Question: Do these quantities for $G(n, r_n)$ converge to analogous quantities of interest for D?

Optimal cuts of a bounded domain $D \subset \mathbb{R}^d$





Define the **minimal bisection** and **Cheeger constant** of D by

$$\mathrm{MBIS}(D) := \inf\{|\partial_D A| : A \subset D, |A| = 1/2\}$$

$$\mathrm{CHE}(D) := \inf\left\{\frac{|\partial_D A|}{|A|} : A \subset D, 0 < |A| \leq |D|/2\right\},$$

$|A|$ denotes the volume of A, $|\partial_D A|$ denotes the perimeter of A within D, i.e. the surface measure of $\overline{A} \cap \overline{D \setminus A}$ (where $\overline{A}$ means closure of A).

[Cheeger's inequality: $\lambda_1 \geq \frac{(\mathrm{CHE}(D))^2}{4}$, where λ_1 is the first non-zero eigenvalue of $-\triangle$ on D.]

Can we learn about D from the sample V_n?

In particular, about $\mathrm{CHE}(D)$ from $\mathrm{CHE}(G(n, r_n))$, given $(r_n)_{n \geq 1}$?

Given $U \subset V_n$, we'll use notation

$$\partial_n(U) := \partial_{G(n,r_n)}(U),$$
$$\mathrm{vol}_n(U) := \mathrm{vol}_{G(n,r_n)}(U) = \#(U)/n.$$

Also, assume that $r_n \ll 1$ and (unless stated otherwise) that

$$nr_n^d \gg \log n,$$

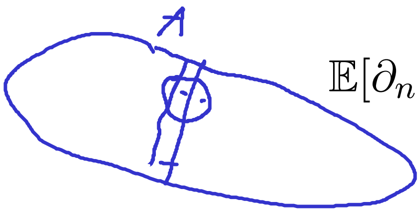
where $a_n \ll b_n$ or $b_n \gg a_n$ means $(a_n/b_n) \to 0$ as $n \to \infty$.

Note: $\exists c > 0$: if $nr_n^d \leq c \log n$ then G is not connected so $\mathrm{CHE}(G) = 0$. Need at least $nr_n^d \geq c \log n$ to have any chance of learning anything from $\mathrm{CHE}(G(n, r_n))$. But want r_n small for computational reasons.

Asymptotic upper bound for $\mathrm{CHE}(G)$

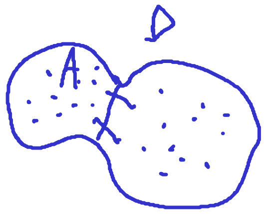
$$\mathrm{CHE}(G(n,r_n)) = \min\left\{\frac{\partial_n(U)}{\mathrm{vol}_n(U)} : U \subset V_n, 0 < \mathrm{vol}_n(U) \leq 1/2\right\}$$

Choose $A \subset D$ to minimize $|\partial_D A|/|A|$ subject to $0 < |A| \leq \frac{1}{2}$.
Let $U_n = V_n \cap A$. By the Law of Large Numbers, $\mathrm{vol}_n(U_n) \to |A|$. Also,



$$\mathbb{E}[\partial_n(U_n)] = n^2 \int_A \int_{D\setminus A} \mathbf{1}_{[0,r_n]}(|y-x|)dydx$$

$$\sim |\partial_D A|\sigma n^2 r_n^{d+1},$$



with $\sigma := (1/2)\int_{\mathbb{R}^d} x_1 \mathbf{1}_{[0,1]}(|x|)dx$. ['Surface tension' of $\phi = \mathbf{1}_{[0,1]}$].
So assuming $\partial_n(U_n) \sim \mathbb{E}[\partial_n(U_n)]$, as $n \to \infty$

$$\limsup n^{-2} r_n^{-d-1} \mathrm{CHE}(G(n,r_n)) \leq \limsup n^{-2} r_n^{-d-1}\left(\frac{\partial_n(U_n)}{\mathrm{vol}_n(U_n)}\right)$$

$$= \frac{\sigma|\partial_D A|}{|A|} = \sigma \mathrm{CHE}(D)$$

Theorem (García Trillos et al. '16; Müller/P. '20)

[Recall $\mathrm{CHE}(D) := \inf\left\{\frac{|\partial_D A|}{|A|} : A \subset D, 0 < |A| \le |D|/2\right\}$
$\mathrm{CHE}(G) = \min\left\{\frac{\partial_G(U)}{\mathrm{vol}_G(U)} : U \subset V(G), 0 < \mathrm{vol}_G(U) \le 1/2\right\}$]

Under our conditions ($|D| = 1$, ∂D Lipschitz, $r_n \to 0$, $nr_n^d \gg \log n$), a.s.:

- $n^{-2} r_n^{-d-1} \mathrm{CHE}(G(n, r_n)) \underline{\to \sigma \mathrm{CHE}(D)}$. [already shown $\le$]
- If $A \subset D$ is the (essentially) unique Cheeger minimizer, i.e. $|A| < 1/2$ and $\frac{|\partial_D A|}{|A|} < \frac{\partial_D A'}{|A'|}$ for all $A' \subset D$ with $|A' \triangle A| \ne 0$, then

  $\forall A_n$ minimizing in $\mathrm{CHE}(G(n,r_n))$

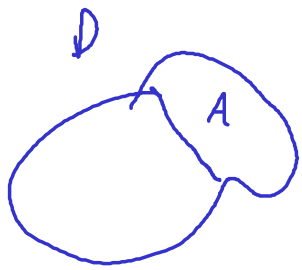
$$n^{-1} \sum_{x \in A_n} \delta_x \to \mathrm{Leb}_d|_A \quad \text{weakly.}$$

- If A is not unique, we still have convergence on a subsequence.
- Also $n^{-2} r_n^{-d-1} \mathrm{MBIS}(G(n, r_n)) \to \sigma \mathrm{MBIS}(D)$,

G. Trillos et al. needed the additional condition $nr_n^2 \gg (\log n)^{3/2}$ if $d = 2$.

Sketch proof of lower bound

- Let $U_n \subset V_n, n \geq 1$ be any sequence of Cheeger minimisers in $G(n, r_n)$. Label points of U_n 'red', points of $V_n \setminus U_n$ 'green'.
- Divide D into cubes (boxes) of side $\gamma_n r_n$, where γ_n is a sequence of constants with $1 \gg \gamma_n$ and $n(\gamma_n r_n)^d \gg \log n$.
- WHP, each box contains about $n(\gamma_n r_n)^d$ points of V_n.
- All the boxes must be 'mostly red' or 'mostly green'.
- Let U_n^* be the union of 'mostly red' boxes. Then



$$n^{-2} r_n^{-d-1} \partial_n(U_n) \approx r_n^{-d-1} \int_D \int_D \phi\left(\frac{|x-y|}{r_n}\right) |\mathbf{1}_{U_n^*}(y) - \mathbf{1}_{U_n^*}(x)| dy dx$$

$=: F_n(\mathbf{1}_{U_n^*})$, where $F_n(\mathbf{1}_B)$ is a smoothed measure of $|\partial B|$, $B \subset D$.
- Homogeneity: $F_n(af) = aF_n(f)$ for all $f \in L^1(D)$ and $a > 0$.

Continuing, recall U_n is a Cheeger minimiser in $G(n, r_n) =: G_n$.

$$n^2 r_n^{-d-1} \mathrm{CHE}(G_n) = n^2 r_n^{-d-1} \frac{\partial_n(U_n)}{\mathrm{vol}_n(U_n)} \approx F_n(g_n)$$

where we define $g_n := |U_n^*|^{-1} \mathbf{1}_{U_n^*} \in L^1(D)$ and for $g \in L^1(D)$ we define

$$F_n(g) := r_n^{-d-1} \int_D \int_D \phi\left(\frac{|y-x|}{r_n}\right) |g(y) - g(x)| dy dx.$$

The g_n are bounded in L^1, and by a compactness result of Garcia Trillos and Slepčev 2016), there exist $g \in L^1(D)$ and a subsequence of $\mathbb{N}$ with $g_n \to g$ in L^1 as n goes to infinity along the subsequence. Then by a Gamma-convergence result (also GT&S 2016),

$$\liminf F_n(g_n) \geq F(g)$$

where $F : L^1(D) \to \mathbb{R}$ is homogeneous and for $A \subset D$, we have

$$F(\mathbf{1}_A) = \sigma |\partial_D A|.$$

But $g = \mathbf{1}_A / |A|$ for some A so $F(g) = |\partial_D A| / |A| \geq \mathrm{CHE}(\overset{D}{\cancel{F}})$. $\square$

The largest component of $G_n := G(n, r_n)$

Let $L(G)$ be the number of vertices in the largest component of G.

The asymptotic behaviour of $L(G_n)$ is governed by that of nr_n^d
[note the average degree $\sim nr_n^d \omega_d$, where $\omega_d =$ vol. of unit ball]

If $\lim_{n\to\infty} nr_n^d < \lambda_c(d)$ then $n^{-1}L(G_n) \xrightarrow{P} 0$ as $n \to \infty$.
If $nr_n^d = \lambda > \lambda_c(d)$ then $n^{-1}L(G_n) \xrightarrow{P} \theta(\lambda) \in (0, \infty)$.

($\lambda_c(d)$ is a percolation threshold, not known explicitly.)
This is called a **giant component** phenomenon.

If $nr_n^d \to \infty$ but $nr_n^d/(\log n) \to 0$, then $n - L(n) = I(G_n)$ to first order.
[P. and Yang, in preparation]

Open problems (now $G_n := G(n, r_n)$)

We know $a_n \mathrm{CHE}(G_n) \to \mathrm{CHE}(D)$ and $a_n \mathrm{MBIS}(G_n) \to \mathrm{MBIS}(D)$ when $1 \gg r_n \gg ((\log n)/n)^{1/d}$.

Can we extend this to when $r_n = c((\log n)/n)^{1/d}$, large c?

Or even to when $r_n \gg n^{-1/d}$, at least for MBIS?
(If $n^{-1/d} \ll r_n \ll ((\log n)/n)^{1/d}$ then $1 \ll (n - L(G_n)) \ll n$, where $L(G)$ is the order of the largest component of G.)

Or to the k-nearest-neighbour graph on V_n where $k = k(n) \gg \log n$?
(connect each vertex by an undirected edge to its k nearest neighbours).

Or to other point processes on D, e.g. a regular grid?

References

- García Trillos, N. and Slepčev, D. (2016) Continuum limit of total variation on point clouds. *Arch. Ration. Mech. Anal.* **220**, 193-241.
- García Trillos, N., Slepčev, D., von Brecht, J., Laurent, T. and Bresson, X. (2016) Consistency of Cheeger and ratio cuts. *Journal of Machine Learning Research*.
- Müller, T. and Penrose, M.D. (2020) Optimal Cheeger cuts and bisections of random geometric graphs. *Annals of Applied Probability.*
- Penrose, M. (2003) *Random Geometric Graphs.* Oxford Uni. Press