4.2 Voronoi cells and regularity partitions

Now we are ready to tie regularity partitions to geometric representations. We define the 2-*neighborhood representation* of a graph G as the map $i \mapsto \mathbf{u}_i$, where $\mathbf{u}_i = A^2\mathbf{e}_i$ is the column of A^2 corresponding to node i (where $A = A_G$ is the adjacency matrix of G). Squaring the matrix seems unnatural, but it is crucial. We define a distance between the nodes, called the 2-*neighborhood distance* (or *similarity distance*), by

$$d(s,t) = \frac{1}{n^2}|\mathbf{u}_s - \mathbf{u}_t|_1.$$

This normalization makes it sure that the distance of any two nodes is at most 1. We need some more notation: For a nonempty set $S \subseteq V$, we consider the average distance from S:

$$\overline{d}(S) = \frac{1}{n}\sum_{i\in V} d(i,S) = \frac{1}{n}\sum_{i\in V}\min_{j\in S} d(i,j).$$

Example 4.2.1 To illustrate the substantial difference between the 1-neighborhood and 2-neighborhood metrics, let us consider a random graph with a very simple structure: Let $V(G) = V_1 \cup V_2$, where $|V_1| = |V_2| = n/2$, and let any node in V_1 be connected to any node in V_2 with probability 1/2. With high probability, the ℓ_1 distance of any two columns of the adjacency matrix is of the order n (approximately $n/2$ for two nodes in different classes, and $n/4$ for two nodes in the same class). But if we square the matrix, the ℓ_1 distance of two columns in different classes will be approximately $n^2/4$, while for two columns in the same class it will be $O(n^{3/2})$. With the normalization above, the two classes will be collapsed to single points (asymptotically, of course), but the distance of these two points will remain constant. So the 2-neighborhood distance reflects the structure of the graph very nicely! ♦

Let V be any set, together with a metric d. We define the *Voronoi partition* induced by a subset $S \subseteq V$ as the partition that has a partition class ("cell") V_s for each $s \in S$, and every point $v \in V$ is put in a the cell V_s for which $s \in S$ is a point of S closest to v. For our purposes, ties can be broken arbitrarily. If the metric space is a euclidean space, then Voronoi cells have many nice geometric properties (for example, they are convex polyhedra; see Figure 4.1 for a picture in two dimensions). In our case the Voronoi cells will not be so nice, but there is no principal difference.

Theorem 4.2.2 *Let G be a simple graph, and let $d(.,.)$ be its 2-neighborhood distance.*

(a) *The Voronoi cells of a nonempty set $S \subseteq V$ define a partition $\mathcal{P}$ of V such that $d_\square(G, G_\mathcal{P}) \le 8\overline{d}(S)^{1/2}$.*

(b) *For every partition $\mathcal{P} = \{V_1,\dots,V_k\}$ we can select elements $s_i \in V_i$ so that $S = \{s_1,\dots,s_k\}$ satisfies $\overline{d}(S) \le 4d_\square(G, G_\mathcal{P})$.*

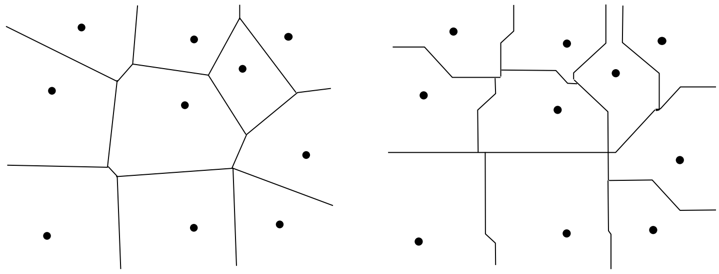

Figure 4.1: Voronoi cells of a finite point set in the plane in Euclidean and Manhattan distance

Proof. In both parts of the proof we work with linear algebra, using the adjacency matrix $A = A_G$. In both parts we consider a particular partition $\mathcal{P} = \{V_1, \dots, V_k\}$. We will be interested in the "error" matrix $R = A - A_{\mathcal{P}} = A - PAP$, for which $\|R\|_\square = d_\square(G, G_{\mathcal{P}})$.

(a) Let $S = \{s_1, \dots, s_k\}$, and let $\mathcal{P}$ be the partition of V defined by the Voronoi cells of S (where $s_i \in V_i$). Recall the definition

$$\|R\|_\square = \frac{1}{n^2} \max_{\mathbf{x},\mathbf{y}\in\{0,1\}^V} |\mathbf{x}^\mathsf{T} R\mathbf{y}|.$$

Let $\mathbf{x}, \mathbf{y}$ be the maximizers on the right, and let $\mathbf{w} = \mathbf{x} - \mathbf{x}_{\mathcal{P}}$ and $\mathbf{z} = \mathbf{y} - \mathbf{y}_{\mathcal{P}}$. The crucial equation is

$$\mathbf{x}^\mathsf{T} R\mathbf{y} = \mathbf{x}^\mathsf{T} A\mathbf{y} - \mathbf{x}^\mathsf{T} A_{\mathcal{P}}\mathbf{y} = \mathbf{x}^\mathsf{T} A\mathbf{y} - \mathbf{x}_{\mathcal{P}}^\mathsf{T} A\mathbf{y}_{\mathcal{P}} = \mathbf{x}^\mathsf{T} A\mathbf{z} + \mathbf{y}_{\mathcal{P}}^\mathsf{T} A\mathbf{w},$$

which implies that

$$|\mathbf{x}^\mathsf{T} R\mathbf{y}| \le |\mathbf{x}|\,|A\mathbf{z}| + |\mathbf{y}_{\mathcal{P}}|\,|A\mathbf{w}| \le \sqrt{n}(|A\mathbf{w}| + |A\mathbf{z}|). \tag{4.1}$$

To estimate $|A\mathbf{z}|$ (say), let $\phi(v) = s_t$ for $v \in V_t$. The fact that we have a Voronoi partition means that $d(v, S) = d(v, \phi(v))$ for every node v. We have

$$A^2\mathbf{z} = \sum_v z_v \mathbf{u}_v = \sum_v z_v(\mathbf{u}_v - \mathbf{u}_{\phi(v)})$$

(since $\sum_{v\in V_t} z_v = 0$). Using that $|z_v| \le 1$ for all $v \in [n]$, we get

$$\begin{aligned} |A\mathbf{z}|^2 &= \mathbf{z}^\mathsf{T}\Big(\sum_v z_v(\mathbf{u}_v - \mathbf{u}_{\phi(v)})\Big) \le \Big|\sum_v z_v(\mathbf{u}_v - \mathbf{u}_{\phi(v)})\Big|_1 \le \sum_v |\mathbf{u}_v - \mathbf{u}_{\phi(v)}|_1 \\ &= n^2 \sum_v d(v, \phi(v)) = n^2 \sum_v d(v, S) = n^3 \overline{d}(S). \end{aligned}$$

We get the same upper bound for $|A\mathbf{w}|$. Combining with (4.1), we get

$$d_\square(G, G_{\mathcal{P}}) = \frac{1}{n^2}|\mathbf{x}^\mathsf{T} R\mathbf{y}| \le \frac{1}{n^{3/2}}(|A\mathbf{w}| + |A\mathbf{z}|) \le 2\sqrt{\overline{d}(S)}.$$

(b) Let i, j be two nodes in the same partition class of $\mathcal{P}$, then $P\mathbf{e}_i = P\mathbf{e}_j$, and hence $A(\mathbf{e}_i - \mathbf{e}_j) = R(\mathbf{e}_i - \mathbf{e}_j)$. Thus

$$d(i, j) = |A^2\mathbf{e}_i - A^2\mathbf{e}_j)|_1 = |AR(\mathbf{e}_i - \mathbf{e}_j)|_1 \le |AR\mathbf{e}_i|_1 + |AR\mathbf{e}_j|_1. \tag{4.2}$$

For every set $V_t \in \mathcal{P}$, choose a point $s_t \in V_t$ for which $|AR\mathbf{e}_i|_1$ is minimized over V_t by $i = s_t$, and let $S = \{s_1, \dots, s_k\}$. The following (somewhat peculiar) inequality relating three matrix norms is not hard to prove:

$$\|AB\|_1 \le 4n\|A\|_\square \|B\|_\infty \qquad (B \in \mathbb{R}^{n\times n}). \tag{4.3}$$

Then using (4.2) and (4.3),

$$\begin{aligned} \overline{d}(S) &\le \frac{1}{n}\sum_{t=1}^{k}\sum_{i\in V_t} d(i, s_t) \le \frac{1}{n^3}\sum_{t=1}^{k}\sum_{i\in V_t}\big(|AR\mathbf{e}_i|_1 + |AR\mathbf{e}_{s_t}|_1\big) \\ &\le \frac{2}{n^3}\sum_{i} |AR\mathbf{e}_i|_1 = \frac{2}{n}\|AR\|_1 \le 4\|R\|_\square = 4d_\square(G, G_{\mathcal{P}}). \end{aligned}$$

□

Combining with the Weak Regularity Lemma, it follows that every graph has an "average representative set" in the following sense.

Corollary 4.2.3 *For every simple graph G and every $k \ge 1$, there is a set $S \subseteq V$ of k nodes such that $\overline{d}(S) \le 16/\sqrt{\log k}$.*