Voronoi Density Estimator for High-Dimensional Data: Computation, Compactification and Convergence (Supplementary Material)

Vladislav Polianskii*[1] **Giovanni Luca Marchetti*[1]** **Alexander Kravberg[1]** **Anastasiia Varava**
Florian T. Pokorny[1] **Danica Kragic[1]**

[1]School of Electrical Engineering and Computer Science, Royal Institute of Technology (KTH), Stockholm, Sweden

We provide here a proof of our main theoretical result with full details.

Theorem 4.1. *Suppose that ρ has support in the whole $\mathbb{R}^n$. For any $K \in L^1(\mathbb{R}^n \times \mathbb{R}^n)$ the sequence of random probability measures $\mathbb{P}_m = f \mathrm{d}x$ defined by the CVDE with m generators converges to $\mathbb{P}$ in distribution w.r.t. x and in probability w.r.t. P. Namely, for any measurable set $E \subseteq \mathbb{R}^n$ the sequence $\mathbb{P}_m(E)$ of random variables over P sampled from ρ converges in probability to the constant $\mathbb{P}(E)$.*

We shall first build up some machinery necessary for the proof. First of all, the following fact on higher-dimensional Euclidean geometry will come in hand.

Proposition 4.2. (Gibbs and Chen 2020, Lemma 5.3) *Let $x \in \mathbb{R}^n$, $\delta > 0$. There exist constants $1 < c_1 < c_2 - 1 < 31$ such that for any open cone $K \subseteq \mathbb{R}^n$ centered at x of solid angle $\frac{\pi}{12}$ and any $p, q, z \in K$, if*

$$d(x,p) < \delta,\ c_1\delta \leq d(x,q) < c_2\delta,\ d(x,z) \geq 32\delta$$

then $d(z,q) < d(z,p)$.

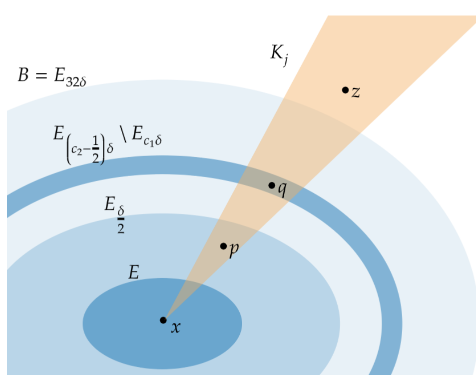

Figure 1: Graphical depiction of sets and points appearing in the proof of Proposition 4.3.

We can now deduce the following.

Proposition 4.3. *Let $\emptyset \neq E \subseteq \mathbb{R}^n$ be a bounded measurable set. There exists a bounded measurable set $B \supseteq E$ such that as $m = |P|$ tends to ∞, the probability with respect to $P \sim \rho^m$ that every Voronoi cell intersecting E is contained in B tends to 1.*

*Equal contribution.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

Proof. Let $\delta = 2\text{diam}\, E = 2\sup_{x,y\in E} d(x,y)$ be twice the diameter of E. For $L > 0$, consider the L-neighbourhood of E

$$E_L = \{x \in X \mid d(x,E) < L\}.$$

First of all, if E has vanishing measure, we can replace it without loss of generality by some E_L, which has nonempty interior.

We claim that $B = E_{32\delta}$ is as desired. To see that, consider an arbitrary $x \in E$ and let $\{K_j\}_j$ be a finite minimal set of open cones centered at x of solid angle $\frac{\pi}{12}$ whose closures cover $\mathbb{R}^n$. As m tends to ∞, since ρ has support in the whole $\mathbb{R}^n$, by the law of large numbers the probability of the following tends to 1:

- P intersects E (recall that E has non-vanishing measure),
- for every j, P intersects $(E_{(c_2-\frac{1}{2})\delta} \setminus E_{c_1\delta}) \cap K_j$, where c_1, c_2 are the constants from Proposition 4.2.

To prove our claim, we can thus conditionally assume the above. Consider now a Voronoi cell intersecting E and suppose by contradiction that z is an element of the cell not contained in B. Let $q \in P$ be a generator in $(E_{(c_2-\frac{1}{2})\delta} \setminus E_{c_1\delta}) \cap K_j$ where K_j is the cone containing z. Since P intersects E, the generator p of the cell lies in $E_{\text{diam}(E)} = E_{\frac{\delta}{2}}$ and consequently $d(x,p) < \delta$. If $p \notin K_j$, then one can replace it with its orthogonal projection on the line passing through x and z. The hypotheses of Proposition 4.2 are then satisfied and we conclude that $d(z,q) < d(z,p)$. This is absurd since p is the generator of $C(z)$. □

For a bounded measurable set $E \subseteq \mathbb{R}^n$, denote by

$$D_E = \max_{\substack{p\in P \\ C(p)\cap E\neq\emptyset}} \text{diam}\, C(p)$$

the maximum diameter of a Voronoi cell intersecting E.

Proposition 4.4. *D_E, thought as a random variable in P, converges in probability to 0 as $m = |P|$ tends to ∞.*

Proof. The proof is inspired by Theorem 4 in Devroye et al. 2015. Consider a finite minimal set of open cones $\{K_j\}_j$ centered at 0 of solid angle $\frac{\pi}{12}$ whose closures cover $\mathbb{R}^n$. Then there is a constant $c > 0$ such that for each $p \in P$

$$\text{diam}\, C(p) \leq c \max_j R_{p,j}$$

where $R_{p,j} = \min_{q\in P\cap(p+K_j)} d(p,q)$ denotes the distance from p to its closest neighbour in the cone K_j centered in p (and $R_{p,j} = \infty$ if $P \cap (p + K_j) = \emptyset$). This follows from Proposition 4.2 applied with $x = p$ to all the cones centered at the generators, with an opportune δ for each of them. For each $\varepsilon > 0$ we thus have an inclusion of events

$$\{D_E > \varepsilon\} \subseteq \left\{ \max_{\substack{p,j \\ C(p)\cap E\neq\emptyset}} R_{p,j} > \frac{\varepsilon}{c} \right\} \subseteq \bigcup_{i,j} \left\{ P \cap (p_i + K_j) \cap B\left(p_i, \frac{\varepsilon}{c}\right) = \emptyset \text{ and } C(p_i) \cap E \neq \emptyset \right\}$$

where $B(x,r)$ is the open ball centered in x of radius r. In the above, we assumed that the set P is equipped with an ordering. For $x \in \mathbb{R}^n$ denote by $E_{x,j}$ the event appearing at the right member of the above expression for $x = p_i$. We can then bound the probability with respect to a random $P \sim \rho^m$, with $m = |P|$ fixed, as

$$\mathbb{P}_{P\sim\rho^m}(D_E > \varepsilon) \leq \sum_{i,j} \mathbb{P}_{P\sim\rho^m}(E_{p_i,j}) = m \sum_j \int_{\mathbb{R}^n} \rho(x) \mathbb{P}_{P\sim\rho^m}(E_{x,j} \mid p_1 = x)\, \mathrm{d}x.$$

Since the points in P are sampled independently we have

$$\mathbb{P}_{P\sim\rho^m}(E_{x,j} \mid p_1 = x,\ C(x) \cap E \neq \emptyset) = \left(1 - \mathbb{P}\left((x + K_j) \cap B\left(x, \frac{\varepsilon}{c}\right)\right)\right)^{m-1} := (1 - M(x))^{m-1}.$$

Pick the set B guaranteed by Proposition 4.3. We can then conditionally assume that every Voronoi cell intersecting E is contained in B, which implies $\mathbb{P}_{P\sim\rho^m}(E_{x,j}) = 0$ for $x \notin B$. The limit we wish to estimate reduces to

$$\lim_{m\to\infty} m \sum_j \int_{\mathbb{R}^n} \rho(x)\mathbb{P}_{P\sim\rho^m}(E_{x,j} \mid p_1 = x)\,\mathrm{d}x = \sum_j \lim_{m\to\infty} \int_B \rho(x) m(1-M(x))^{m-1}\,\mathrm{d}x.$$

Since B is bounded and ρ has support in the whole $\mathbb{R}^n$, $M(x)$ is (essentially) bounded from below by a strictly positive constant as x varies in B. The limit can thus be brought under the integral and putting everything together we get:

$$\lim_{m\to\infty} \mathbb{P}_{P\sim\rho^m}(D_E > \varepsilon) \le \sum_j \int_B \rho(x) \lim_{m\to\infty} m(1-M(x))^{m-1}\,\mathrm{d}x = 0.$$

□

We are now ready to prove Theorem 4.1.

Proof. By the Portmanteau Lemma (Van der Vaart 2000), it is sufficient to that $\mathbb{P}_m(E)$ converges to $\mathbb{P}(E)$ in probability for any bounded measurable set $E \subseteq \mathbb{R}^n$ which is a continuity set for $\mathbb{P}$ i.e., $\mathbb{P}(\partial E) = 0$ where ∂E is the (topological) boundary of E. Pick such E. By definition of the CVDE, for a fixed set P of generators we have that

$$\begin{aligned}\mathbb{P}_m(E) &= \frac{1}{m}|\{p \in P \mid C(p) \subseteq E\}| + \overbrace{\frac{1}{m} \sum_{\substack{p\in P \\ C(p)\not\subseteq E \\ C(p)\cap E\neq\emptyset}} \frac{\mathrm{Vol}_p(C(p)\cap E)}{\mathrm{Vol}_p(C(p))}}^{\overline{R}} \\ &= \frac{1}{m}|P\cap E| + \overline{R} - \frac{1}{m}|\{p \in P\cap E \mid C(p) \not\subseteq E\}|.\end{aligned} \tag{1}$$

Since the Voronoi cells are closed, any cell intersecting E not contained in E intersects ∂E. Thus $\left|\overline{R} - \frac{1}{m}|\{p \in P\cap E \mid C(p) \not\subseteq E\}|\right| \le 2R$ where $R := \frac{1}{m}|\{p \in P \mid C(p)\cap \partial E \neq \emptyset\}|$. Now, the random variable $\frac{1}{m}|P\cap E|$ tends to $\mathbb{P}(E)$ in probability as m tends to ∞ by the law of large numbers. In order to conclude, we need to show that R tends to 0 in probability.

Fix $\varepsilon > 0$. For $L > 0$, consider the L-neighbour $\partial E_L = \{x \in X \mid d(x, \partial E) < L\}$ of the boundary ∂E. If the diameter of the Voronoi cells intersecting ∂E is less than L then all such cells are contained in ∂E_L. Thus:

$$\begin{aligned}\mathbb{P}_{P\sim\rho^m}(R > \varepsilon) &\le \mathbb{P}_{P\sim\rho^m}\left(\frac{1}{m}|P\cap\partial E_L| > \varepsilon \text{ and } D_{\partial E} < L\right) + \mathbb{P}_{P\sim\rho^m}(D_{\partial E} \ge L) \\ &\le \mathbb{P}_{P\sim\rho^m}\left(\frac{1}{m}|P\cap\partial E_L| > \varepsilon\right) + \mathbb{P}_{P\sim\rho^m}(D_{\partial E} \ge L) \\ &\le \mathbb{P}_{P\sim\rho^m}\left(\left|\mathbb{P}(\partial E_L) - \frac{1}{m}|P\cap\partial E_L|\right| > \varepsilon - \mathbb{P}(\partial E_L)\right) + \mathbb{P}_{P\sim\rho^m}(D_{\partial E} \ge L).\end{aligned} \tag{2}$$

Since ∂E is closed, $\partial E = \cap_{L>0}\partial E_L$ and thus $\lim_{L\to 0}\mathbb{P}(\partial E_L) = \mathbb{P}(\cap_L \partial E_L) = \mathbb{P}(\partial E) = 0$ since E is a continuity set. This implies that there is an L such that $\varepsilon > \mathbb{P}(\partial E_L)$. The right hand side of Equation 2 tends then to 0 by the law of large numbers and Proposition 4.4, which concludes the proof.

□

REFERENCES

Devroye, Luc et al. (Dec. 2015). "On the measure of Voronoi cells." In: *Journal of Applied Probability* 54. DOI: `10.1017/jpr.2017.7`.

Gibbs, Isaac and Linan Chen (2020). "Asymptotic properties of random Voronoi cells with arbitrary underlying density." In: *Advances in Applied Probability* 52.2, pp. 655–680.

Van der Vaart, Aad W (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.