An Efficient and Continuous Voronoi Density Estimator

Giovanni Luca Marchetti   **Vladislav Polianskii**   **Anastasiia Varava**
Florian T. Pokorny   **Danica Kragic**
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology
Stockholm, Sweden

Abstract

We introduce a non-parametric density estimator deemed Radial Voronoi Density Estimator (RVDE). RVDE is grounded in the geometry of Voronoi tessellations and as such benefits from local geometric adaptiveness and broad convergence properties. Due to its radial definition RVDE is continuous and computable in linear time with respect to the dataset size. This amends for the main shortcomings of previously studied VDEs, which are highly discontinuous and computationally expensive. We provide a theoretical study of the modes of RVDE as well as an empirical investigation of its performance on high-dimensional data. Results show that RVDE outperforms other non-parametric density estimators, including recently introduced VDEs.

1 INTRODUCTION

The problem of estimating a Probability Density Function (PDF) from a finite set of samples lies at the heart of statistics and arises in several practical scenarios (Diggle 2013; Scott 2015). Among density estimators, the non-parametric ones aim to infer a PDF through a closed formula. Differently from parametric methods, they do not require optimization and ideally provide an estimated PDF which is simple, interpretable and computationally efficient. Two traditional examples of non-parametric density estimators are the Kernel Density Estimator (KDE; Gramacki 2018, Rosenblatt 1956) and histograms (Freedman et al. 1981, Pearson 1894). KDE consists of a mixture of local copies of a kernel around each datapoint while histograms partition the ambient space into local cells ('bins') where the estimated PDF is constant.
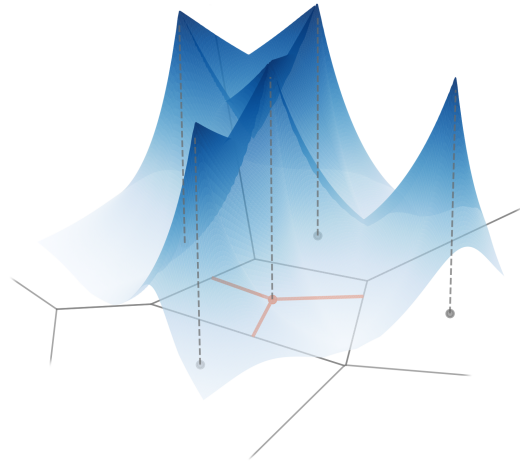

Figure 1: An example of a density estimated via RVDE. The Voronoi tessellation is depicted in solid gray. The estimated density is defined by the property that its conical integral over the rays originating from the datapoints (orange) is constant.

Both histograms and KDE suffer from bias due to the prior choice of a local geometric structure i.e., the bins and the kernel respectively. This bias gets exacerbated in high-dimensional ambient spaces. The reason is that datasets grow exponentially in terms of geometric complexity, making a fixed simple geometry unsuitable for estimating high-dimensional densities. This has led to the introduction of the *Voronoi Density Estimator* (VDE; Ord 1978). VDE relies on the geometric adaptiveness of Voronoi cells, which are convex polytopes defined locally by the data (Okabe et al. 2009). The PDF estimated by VDE is constant on such cells, thus behaving as an adaptive version of histograms. Due to its local geometric properties, VDE possesses convergence guarantees to the ground-truth PDF which are more general than the ones of KDE.

The geometric benefits of VDE, however, come with a number of shortcomings. First, the Voronoi cells and in particular their volumes are computationally expensive to

compute in high dimensions. Although this has been recently attenuated by proposing Monte Carlo approximations (Polianskii et al. 2022), VDE falls behind methods such as KDE in terms of computational complexity. Second, VDE (together with its generalized version from Polianskii et al. 2022 deemed CVDE) is highly discontinuous on the boundaries of Voronoi cells. The estimated PDF consequently suffers from large variance and instability with respect to the dataset. This is again in contrast to KDE, which is continuous in its ambient space.

In this work, we propose a novel non-parametric density estimator deemed *Radial Voronoi Density Estimator* (RVDE) which addresses the above challenges. Similarly to VDE, RVDE integrates to a constant on Voronoi cells and thus shares its local geometric advantages and convergence properties. In contrast to VDE, RVDE is continuous and computable in linear time with respect to the dataset size. The central idea behind RVDE is to define the PDF radially from the datapoints so that the (conic) integral over the ray cast in the corresponding Voronoi cell is constant (see Figure 1). This is achieved via a 'radial bandwidth' which is defined implicitly by an integral equation. Intuitively, the radial approach reduces the high-dimensional geometric challenge of defining a Voronoi-based estimator to a one-dimensional problem. This avoids the expensive volume computations of the original VDE and guarantees continuity because of the fundamental properties of Voronoi tessellations. Another important aspect of RVDE is its geometric distribution of modes. We show that the modes either coincide with the datapoints or lie along the edges of the Gabriel graph (Gabriel et al. 1969) depending on a hyperparameter analogous to the bandwidth in KDE.

We compare RVDE with CVDE, KDE and the adaptive version of the latter in a series of experiments. RVDE outperforms the baselines in terms of the quality of the estimated density on a variety of datasets. Moreover, it runs significantly faster and with lower sampling variance compared to CVDE. This empirically confirms that the geometric and continuity properties of RVDE translate into benefits for the estimated density in a computationally efficient manner. We provide an implementation of RVDE (together with baselines and experiments) in $C++$ at a publicly available repository [1]. The code is parallelized via the OpenCL framework and comes with a Python interface. In summary our contributions include:

- A novel density estimator (RVDE) based on the geometry of Voronoi tessellations which is continuous and computationally efficient.

- A complete study of the modes of RVDE and their geometric distribution.

- An empirical investigation comparing RVDE to KDE (together with its adaptive version) and previously studied VDEs.

2 RELATED WORK

2.1 Non-parametric Density Estimation

Non-parametric methods for density estimation trace back to the introduction of histograms (Pearson 1894). Histograms have been extended by considering bin geometries beyond the canonical rectangular one, for example triangular (Scott 1988) and hexagonal (Carr et al. 1992) geometries. Another popular density estimator is KDE, first discussed by Rosenblatt 1956 and Parzen 1962. The estimated density is a mixture of copies of a priorly chosen distribution ('kernel') centered at the datapoints. KDE has been extended to the multivariate case (Izenman 1991; Dehnad 1987) and has seen improvements such as bandwidth selection methods (Marron 1987; Wand et al. 1994) and algorithms for adaptive bandwidths (Wang et al. 2007; Walt et al. 2017). Applications of KDE include estimation of traffic incidents (Xie et al. 2008), of archaeological data (Baxter et al. 1997) and of wind speed (Bo et al. 2017) to name a few. As discussed in Section 1, both KDE and histograms suffer from lack of geometric adaptiveness due to the choice of prior local geometries.

Another class of non-parametric methods are the orthogonal density estimators (Vannucci 1995; Masry 1997). Those consist of choosing a discretized orthonormal basis of functions and computing the coefficients of the ground-truth density via Monte-Carlo integration over the dataset. When the basis is the Fourier one, the estimator is referred to as 'wavelet estimator'. The core drawback is that orthogonal density estimators do not scale efficiently to higher dimensions. When considering canonical tensor product bases the complexity grows exponentially w.r.t. the dimensionality (Walter 1995), making the estimator unfeasible to compute.

2.2 Voronoi Density Estimators

The first Voronoi Density Estimator (VDE) has been pioneered by Ord 1978. The estimated density relies on Voronoi tessellations in order to achieve local geometric adaptiveness. This is the main advantage of VDE over methods such as KDE. The original VDE has seen applications to real-world densities such as neurons in the brain (Duyckaerts et al. 1994), photons (Ebeling et al. 1993) and stars in a galaxy (Vavilova et al. 2021). However, the method is not immediately extendable to high-dimensional spaces because of unfeasible computational complexity of volumes and abundance of unbounded Voronoi cells. This has been only recently amended by Polianskii et al. 2022 by introducing approximate numerical algorithms and by

---

[1] https://github.com/giovanni-marchetti/rvde

shaping of the density via a kernel. In the present work, we aim to design an alternative version of the original VDE which is continuous and does not rely on volume computations. The resulting estimator is thus stable and computationally efficient while still benefiting from the geometry of Voronoi tessellations.

3 BACKGROUND

In this section we recall the class of non-parametric density estimators which we will be interested in throughout the present work. To this end, let $P \subseteq \mathbb{R}^n$ be a finite set and consider the following central notion from computational geometry.

Definition 1. The *Voronoi cell*[2] of $p \in P$ is:

$$C(p) = \{x \in \mathbb{R}^n \mid \forall q \in P \ d(x,q) \geq d(x,p)\}. \quad (1)$$

The Voronoi cells are convex polytopes that intersect at the boundary and cover the ambient space $\mathbb{R}^n$. The collection $\{C(p)\}_{p \in P}$ is referred to as *Voronoi tessellation* generated by P. Note that although the Voronoi tessellations are defined in an arbitrary metric space, the resulting cells might be non-convex for distances different from the Euclidean one. Since convexity will be crucial for the following constructions, we stick to the Euclidean metric for the rest of the work.

We call *density estimator* any mapping associating a probability density function $f_P \in L^1(\mathbb{R}^n)$ to a finite set $P \subseteq \mathbb{R}^n$. The following class of density estimators generalizes the original one by Ord 1978.

Definition 2. A *Voronoi Density Estimator* (VDE) is a density estimator $P \mapsto f_P$ such that for each $p \in P$:

$$\int_{C(p)} f_P(x) \, \mathrm{d}x = \frac{1}{|P|}. \quad (2)$$

VDEs stand out among density estimators for their geometric properties. This is because the Voronoi cells are arbitrary polytopes that are adapted to the local geometry of data. For VDEs all the Voronoi cells have the same estimated probability, making such estimators locally adaptive from a geometric perspective. This is reflected, for example, by the general convergence properties of VDEs. The following is the main theoretical result from Polianskii et al. 2022.

Theorem 3.1. *Let $P \mapsto f_P$ be a VDE and suppose that P is sampled from a probability density $\rho \in L^1(\mathbb{R}^n)$ with support in the whole $\mathbb{R}^n$. For P of cardinality m consider the probability measure $\mathbb{P}_m = f_P \mathrm{d}x$ which is random in P. Then the sequence $\mathbb{P}_m$ converges to $\mathbb{P} = \rho \mathrm{d}x$ in distribution w.r.t. x and in probability w.r.t. P. Namely, for any*

---

[2]Sometimes referred to as *Dirichlet cell*.

measurable set $E \subseteq \mathbb{R}^n$ the sequence of random variables $\mathbb{P}_m(E)$ converges in probability to the constant $\mathbb{P}(E)$.

In contrast, the convergence of other density estimators such as KDE requires the kernel bandwidth to vanish asymptotically (Devroye et al. 1979). The bandwidth vanishing is necessary in order to amend for the local geometric bias inherent in KDE as discussed in Section 1.

The following canonical construction of a VDE deemed Compactified Voronoi Density Estimator (CVDE) is discussed by Polianskii et al. 2022. Given an integrable kernel $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{>0}$ the estimated density is defined as

$$f_P(x) = \frac{K(p,x)}{|P| \operatorname{Vol}_p(C(p))} \quad (3)$$

where p is the closest point in P to x and $\operatorname{Vol}_p(C(p)) = \int_{C(p)} K(p,y) \, \mathrm{d}y$. The latter volumes are approximated via Monte Carlo methods since they become unfeasible to compute exactly as dimensions grow. The resulting density inherits the same regularity as K when restricted to each Voronoi cell but jumps *discontinuously* when crossing the boundary of Voronoi cells (see Figure 3). Motivated by this, the goal of the present work is to introduce a continuous and efficient VDE.

4 METHOD

4.1 Radial Voronoi Density Estimator

In this section we outline a general way of constructing a VDE with continuous density function. Our central idea is to define the latter *radially* w.r.t. the datapoints $p \in P$. We start by rephrasing the integral over a Voronoi cell (Equation 2) in spherical coordinates:

$$\int_{C(p)} f_P(x) \, \mathrm{d}x =$$
$$= \int_{\mathbb{S}^{n-1}} \underbrace{\int_0^{l(p+\sigma)} t^{n-1} f_P(p+t\sigma) \, \mathrm{d}t}_{\text{Conical Integral}} \mathrm{d}\sigma. \quad (4)$$

Here $\mathbb{S}^{n-1} \subseteq \mathbb{R}^n$ denotes the unit sphere and $l(x) \in [0,+\infty]$ denotes the length of the segment contained in $C(p)$ of the ray cast from p passing through x i.e.,

$$l(x) = \sup \left\{ t \geq 0 \mid p + t \frac{x-p}{d(x,p)} \in C(p) \right\}. \quad (5)$$

We refer to Figure 2 for a visual illustration. Note that $l(x)$ is defined for $x \neq p$ and is continuous in its domain since $l(x) = d(x,p) = d(x,q)$ for $x \in C(p) \cap C(q)$.

We aim to solve Equation 4 by forcing the conical integral in Equation 4 to be *constant*. To this end, we fix a continuous and strictly decreasing function $K : \mathbb{R}_{>A} \to \mathbb{R}_{\geq 0}$
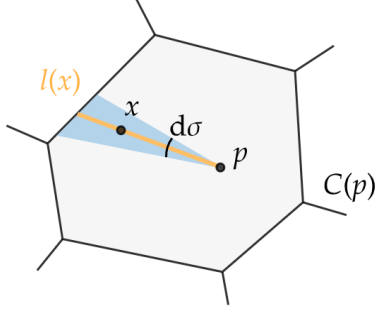
Figure 2: Illustration of the quantity $l(x)$ involved in the definition of RVDE (Definition 3). The estimated density integrates to a constant over all the infinitesimal cones (blue) in $C(p)$ originating from p.

(a 'kernel') defined on a half-line $\mathbb{R}_{>A}$, $A < 0$, with the property that $t^{n-1}K(t)$ is integrable on $\mathbb{R}_{>0}$. By an ansatz we look for a density in the form

$$f_P(x) = \frac{K\left(\beta(l(x))d(x,p)\right)}{\alpha|P|\text{Vol}(\mathbb{S}^{n-1})} \qquad (6)$$

where $\alpha > 0$ is a hyperparameter and $\beta : \mathbb{R}_{>0} \to \mathbb{R}$ is a function that we would like to determine. The latter intuitively represents a *radial bandwidth*. The density f_P is continuous since the discontinuity of l at $x = p$ is amended by the vanishing of $d(x,p)$. Equation 2 is satisfied if for every $l > 0$:

$$\int_0^l t^{n-1}K(\beta(l)t)\,\mathrm{d}t = \alpha. \qquad (7)$$

Since K is strictly decreasing, the above expression always has a unique solution $\beta(l) > \frac{A}{l}$ assuming that $t^{n-1}K(t)$ is not integrable around A. Such a guaranteed solution can be computed via any root-finding algorithm and is continuous w.r.t. l. We provide an analysis of the function β and a discussion of the Newton-Raphson method for its computation in Section 4.3.

The derivations above bring us to the following definition.

Definition 3. Fix an $\alpha > 0$ and a continuous function $K : \mathbb{R}_{>A} \to \mathbb{R}_{>0}$ with the domain bound $A < 0$. Assume the following:

- $K(0) = 1$,

- K is strictly decreasing,

- $|t|^{n-1}K(t)$ is integrable around $+\infty$ but not integrable around A.

The *Radial Voronoi Density Estimator* (RVDE) is the density estimator defined by Equation 6 where β is the function defined implicitly by Equation 7.

The following two standard families of kernels K satisfy the above requirements:

$$\begin{array}{cc} \textit{Exponential} & \textit{Rational} \\ K(t) = e^{-t} & K(t) = \frac{1}{(t+1)^k} \end{array} \qquad (8)$$

where $k > n$. The domain bounds are $A = -\infty$ and $A = -1$ respectively. When $n = 1$ and K is the exponential kernel, the function β is closely related to the Lambert W function (Corless et al. 1996) via the expression:

$$\beta(l) = \frac{1}{\alpha} + W\left(-\frac{l}{\alpha}e^{-\frac{l}{\alpha}}\right). \qquad (9)$$

We provide an empirical comparison between the two kernels from Equation 8 in Section 5.3.

The intuition behind the hyperparameter α is that it controls the trade-off between the amount of density concentrated around P and away from it (i.e., on the boundary of Voronoi cells). Indeed as $\alpha \to 0^+$ RVDE tends (in distribution) to the discrete empirical measure over P while as $\alpha \to +\infty$ it tends to a measure concentrated on the boundary of Voronoi cells. This can be deduced from Equation 7 since $\beta(l)$ tends to $+\infty$ and to A/l respectively and thus Equation 6 tends to 0 for $d(x,p) \neq 0, l(x)$. This intuition around α will be corroborated by Proposition 4.3, where we study how it controls the distribution of modes of RVDE and consequently propose a heuristic selection procedure.

4.2 Computational Complexity and Sampling

We now discuss the computational cost of evaluating RVDE at a point $x \in \mathbb{R}^n$. To begin with, the closest $p \in P$ to x can be found in logarithmic time w.r.t. $|P|$ by organizing P in an efficient data structure for nearest neighbor lookups such as a k-d tree. Then $l(x)$ can be computed in linear time via the following closed expression (Polianskii et al. 2022):

$$l(x) = \min_{q \neq p,\ l^q(x) \geq 0} l^q(x) \qquad (10)$$

where

$$l^q(x) = \frac{d(q,p)^2}{2\left\langle \frac{x-p}{d(x,p)}, q-p \right\rangle}. \qquad (11)$$

The computational cost of evaluating $f_P(x)$ is thus linear w.r.t. $|P|$. The remaining compute essentially reduces to solving Equation 7, which depends on the integrator, the root-finder algorithm adopted and the desired precision.

The formulation of RVDE enables a simple and efficient procedure for sampling from the estimated density. In order to sample, one first chooses a $p \in P$ uniformly since f_P integrates to $\frac{1}{|P|}$ on each Voronoi cell (Equation 2). Since $t^{n-1}f_P(p+t\sigma)$ integrates to a constant on the ray $r = \{p+t\sigma\}_{t \geq 0} \cap C(p)$ for every $\sigma \in \mathbb{S}^{n-1}$, one then

KDE
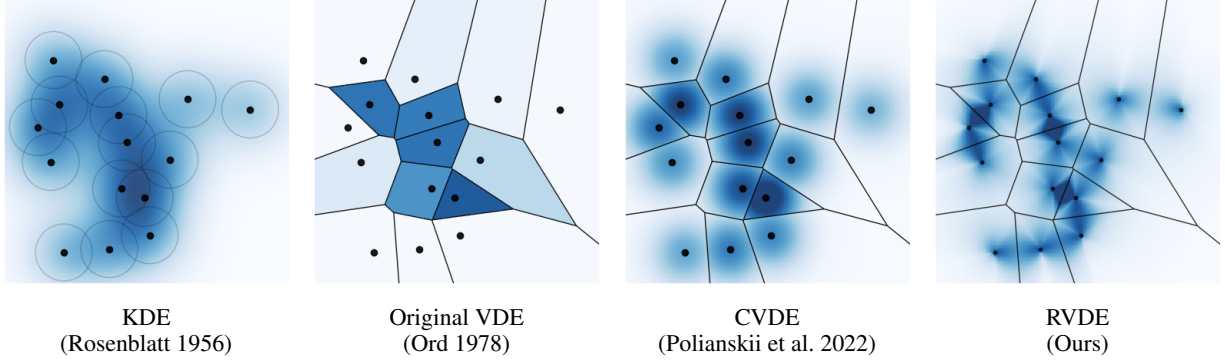(Rosenblatt 1956)

Original VDE
(Ord 1978)

CVDE
(Polianskii et al. 2022)

RVDE
(Ours)

Figure 3: From left to right: heatmaps of KDE, of the two VDEs from the literature and of our RVDE.

samples σ uniformly from the sphere. Finally one samples t from the one-dimensional density $t^{n-1}K(t)$ restricted to the interval $[0, l(p + \sigma)]$. The computational complexity of the latter step depends on the kernel as well as of the sampling method. The result of the sampling is $p + t\sigma$. Because of the computational cost of $l(p + \sigma)$, the sampling complexity of RVDE is linear w.r.t. $|P|$.

RVDE is more efficient than the VDE discussed by Polianskii et al. 2022 (see the end of Section 3). The latter relies on Monte Carlo integration for numerical approximation of volumes of Voronoi cells and has complexity $O(\Sigma|P|^2)$ where Σ is the number of Monte Carlo samples. Compared to KDE, RVDE has the same computational complexity (for both evaluation and sampling) while retaining the geometric benefits of a VDE.

4.3 Study of β and Modes

In this section we discuss qualitative properties and computational aspects of the function β defined implicitly by Equation 7 and consequently characterize the modes of RVDE. We start by presenting an explicit expression of the Newton-Raphson iteration for the computation of $\beta(l)$.

Proposition 4.1. *Fix $l > 0$ and suppose $K \in C^1(\mathbb{R}_{>A})$ i.e., it is continuously differentiable. Then the iteration β_{m+1} of the Newton-Raphson method for computing $\beta(l)$ by solving Equation 7 takes form:*

$$\beta_{m+1} = \tag{12}$$

$$= \beta_m + \frac{\beta_m}{n}\left(1 - \frac{l^n K(\beta_m l) - n\alpha}{l^n K(\beta_m l) - n\int_0^l t^{n-1}K(\beta_m t)\,\mathrm{d}t}\right).$$

Moreover, if K is convex then the Newton-Raphson method converges for any initial value β_0 i.e., $\lim_{m\to\infty}\beta_m = \beta(l)$.

We refer to the Appendix for a proof. Note that the convexity assumption is satisfied by both the kernels from Equation 8. Proposition 4.1 enables to compute $\beta(l)$ and together with Section 4.2 provides all the algorithmic details for implementing RVDE.

Next, we outline a qualitative study of the function $l \mapsto \beta(l)$.

Proposition 4.2. *The function $\beta : \mathbb{R}_{>0} \to \mathbb{R}$ is increasing, has a zero at $l = (n\alpha)^{\frac{1}{n}}$ and has an horizontal asymptote:*

$$\lim_{l\to+\infty}\beta(l) = \left(\frac{1}{\alpha}\int_0^\infty t^{n-1}K(t)\,\mathrm{d}t\right)^{\frac{1}{n}}. \tag{13}$$

Moreover if $K \in C^1(\mathbb{R}_{>A})$ then $\beta \in C^1(\mathbb{R}_{>0})$ and it satisfies the differential equation:

$$\left(l - \frac{n\alpha}{l^{n-1}K(\beta(l)l)}\right)\frac{\mathrm{d}\beta}{\mathrm{d}l}(l) = -\beta(l). \tag{14}$$

We refer to the Appendix for a proof. As discussed in Section 4.1, β generalizes the Lambert W function. The properties and the differential equation from Proposition 4.2 generalize their well-known instances for the W function (Corless et al. 1996).

We now focus on the study of modes. Our goal is to describe the modes of RVDE completely. This is an advantage over density estimators such as KDE, where the modes are challenging to describe and to compute approximately (Lee et al. 2021; Comaniciu et al. 2002). Denote by $\varepsilon = (n\alpha)^{\frac{1}{n}}$ the zero of β. Proposition 4.2 implies that for $x \in \mathbb{R}^n$, the density f_P decreases radially w.r.t. p in the direction of x if $l(x) > \varepsilon$ and increases otherwise. This leads to the following result.

Proposition 4.3. *The modes of f_P are classified as follows:*

(1) $p \in P$ if $d(p, q) > 2\varepsilon$ for every Voronoi cell $C(q)$ adjacent to $C(p)$,

(2) $\frac{p+q}{2}$ for $p, q \in P$ if $\frac{p+q}{2} \in C(p) \cap C(q)$ and $d(p, q) < 2\varepsilon$,

(3) all points belonging to the segment $[p, q]$ for $p, q \in P$ if $\frac{p+q}{2} \in C(p) \cap C(q)$ and $d(p, q) = 2\varepsilon$.

We refer to the Appendix for a proof and to Figure 5 for an illustration. Since ε depends monotonically on the hyperparameter α, the latter controls the threshold for distances

between adjacent points in P below which the mode gets pushed away from such points towards the boundary of the Voronoi cells. Intuitively, α determines the extent by which points in P are considered 'isolated' (i.e., a mode) or otherwise get 'merged' by placing a mode between them.

An alternative geometric formulation of Proposition 4.3 is the following. Consider the *Gabriel graph* of P (Gabriel et al. 1969) containing an edge between p and q iff $\frac{p+q}{2} \in C(p) \cap C(q)$ and discard all the edges of length greater than 2ε. The modes of RVDE are then associated with (1) all isolated vertices, (2) midpoints of edges and (3) whole edges of length 2ε. Intuitively, the modes of RVDE are distributed geometrically according to the truncated Gabriel graph.
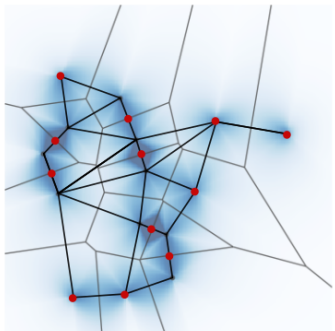

Figure 5: Illustration of the modes of RVDE (red) together with the Gabriel graph (black). Datapoints that not modes are marked in yellow.

This suggests a possible heuristic procedure for *hyperparameter selection* of α. An option is to consider statistics of lengths of edges in the Gabriel graph and choose 2ε (and thus α) as a percentile. The percentile we suggest is $\frac{|P|-1}{|E|}$ where E denotes the set of edges of the Gabriel graph. The intuition is that we wish to avoid modes distributed in cycles. The number of cycles in the Gabriel graph is $|P| - |E| + 1$, from which our suggested percentile follows. This procedure enables to select α automatically and we evaluate it empirically in Section 5. However, it comes with a number of limitations. First, the computational complexity of such a procedure is $O(|P|^3)$ because of the construction of the Gabriel graph, which is feasible but might become expensive for large datasets. Another limitation is its independence from the kernel K. The selection of α might be satisfying for some kernels but not for others. In our empirical evaluation from Section 5 we show that for the rational kernel the selected α is close to the optimal one in practice, while for the exponential kernel the selection is further from optimality.

5 EXPERIMENTS

Our empirical investigation is organized as follows. First we study RVDE on its own by comparing the different choices of the kernel. We then compare RVDE with other non-parametric density estimators on a variety of datasets.

5.1 Evaluation Metrics and Baselines

We evaluate all the density estimators f_P via average log-likelihood on a test set P_{test} i.e.,

$$\frac{1}{|P_{\text{test}}|} \sum_{x \in P_{\text{test}}} \log f_P(x). \tag{15}$$

This measures whether the estimator assigns high density values to points outside of P but sampled from the same distribution. In order to empirically evaluate the computational complexity, we additionally include runtimes for all the experiments. Our implementations of all the considered density estimators share the same programming framework and are parallelized to a similar degree, making the raw runtimes a fair comparison. We perform experiments on a machine with an AMD Ryzen 9 5950X 16-core CPU and a GeForce RTX 3090 GPU.

We deploy the following non-parametric density estimators as baselines in the experiments.

Kernel Density Estimator (KDE): given a (normalized) kernel $K : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ the density is estimated as:

$$f_P(x) = \frac{1}{|P|h^n} \sum_{p \in P} K\left(\frac{x-p}{h}\right) \tag{16}$$

where h is the bandwidth hyperparameter.

Adaptive Kernel Density Estimator (AdaKDE; Wang et al. 2007): a version of KDE where the bandwidth h_p depends on $p \in P$ and is smaller when data is denser around p. Specifically, if $f_P(p)$ denotes the standard KDE estimate with a global bandwidth h then $h_p = h\lambda_p$, where:

$$\lambda_p = (g/f_P(p))^{\frac{1}{2}}, \qquad g = \prod_{q \in P} f_P(q)^{\frac{1}{|P|}}. \tag{17}$$

Compactified Voronoi Density Estimator (CVDE; Polianskii et al. 2022): the VDE described at the end of Section 3. It depends on a kernel K (together with a bandwidth) and is discontinuous on the boundary of Voronoi cells.

5.2 Datasets

In our experiments we consider data of varying nature. This includes both simple synthetic distributions and real-world datasets in high dimensions. For the latter, we consider
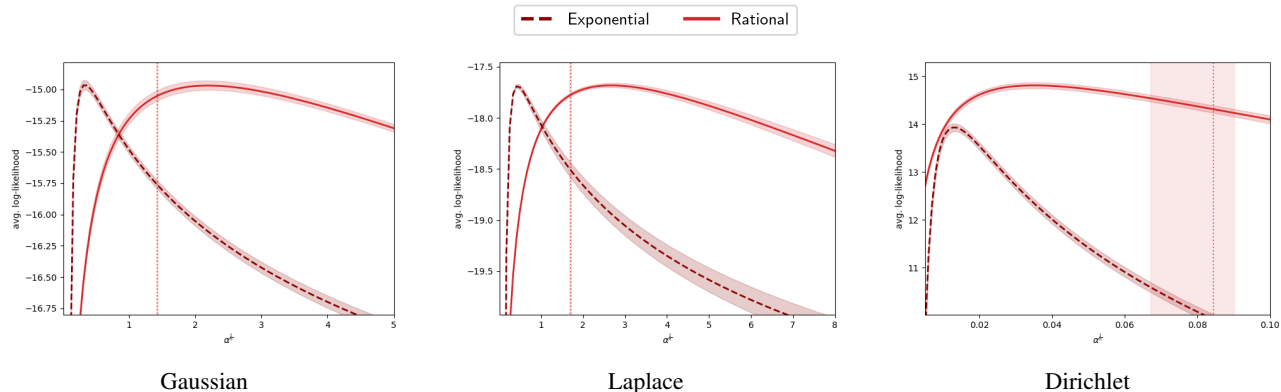
Figure 4: Comparison of the two kernels for RVDE (Equation 8) on three simple distributions in 10 dimensions.

sound data ($n = 21$) and image data ($n = 100$). Our datasets are the following.

Synthetic Datasets: datasets generated from a number of simple densities in $n = 10$ dimensions. Both P and P_{test} contain 1000 points in all the cases. The densities we consider are: a standard Gaussian distribution, a standard Laplace distribution, a Dirichlet distribution with parameters $\alpha_i = \frac{1}{n+1}$ and a mixture of two Gaussians with means $\mu_1 = (-0.5, 0, \cdots, 0)$, $\mu_2 = (0.5, 0, \cdots, 0)$ and standard deviations $\sigma_1 = 0.1$, $\sigma_2 = 10$ respectively.

MNIST (Deng 2012): a dataset consisting of 28×28 grayscale images of handwritten digits which are normalized in order to lie in $[0,1]^{28 \times 28}$. In order to densify the data and obtain more meaningful estimates, we downscale the images to resolution 10×10. For each experimental run, we sample half of the 60000 training datapoints in order to evaluate the variance of the estimation. The test set size is 10000.

Anuran Calls (Dua et al. 2017): a dataset consisting of 7195 calls from 10 species of frogs which are represented by 21 normalized mel-frequency cepstral coefficients in $[0,1]^{21}$. We retain 10% of data for testing and sample half of the training data at each experimental run.

5.3 Comparison of Kernels

Our first experiment consists of a comparison between the rational and the exponential kernel for RVDE (Equation 8) on the synthetic datasets. In what follows the exponent k for the rational kernel is set to $k = n + 1$ for simplicity, where n is the dimension of the ambient space of the dataset considered (in this experiment, $n = 10$).

The results are presented in Figure 4. The plot displays the test log-likelihood (Equation 15) as the hyperparameter α varies. The latter is scaled as $\alpha^{\frac{1}{n}}$ in order to be consistent with the visualizations in the following section. The curves on the plot represent mean and standard de-

viation (shaded areas) over 5 experimental runs for 100 bandwidths. The additional vertical lines correspond to the value of the hyperparameter selection heuristic discussed at the end of Section 4.3. As can be seen, the performance of the rational kernel is more stable w.r.t. the hyperparameter α. The exponential kernel, however, achieves a slightly higher test score with its best hyperparameter on the Gaussian and Laplace datasets. Note that the heuristically chosen α aligns well with the one that achieves the best performance for the rational kernel, but is misaligned for the exponential one. We conclude that the rational kernel is generally a better option unless an extensive hyperparameter search is performed. In what follows we consequently stick to the rational kernel for RVDE.

5.4 Comparison with Baselines

In our main experiment we compare the performance of RVDE with the baselines described in Section 5.1. We consider the test log-likelihood (Equation 5.1), the standard deviation of the latter and the runtimes. In order to make the comparison as fair as possible, all the estimators are implemented with the rational kernel. We found out that the performances drop with the more standard Gaussian kernel (which does not apply to RVDE). We include the results with both the Gaussian kernel and the exponential kernel in the Appendix.

Table 1: Average runtimes (in seconds) per one full train-test run with fixed bandwidth. RVDE is highlighted in blue.

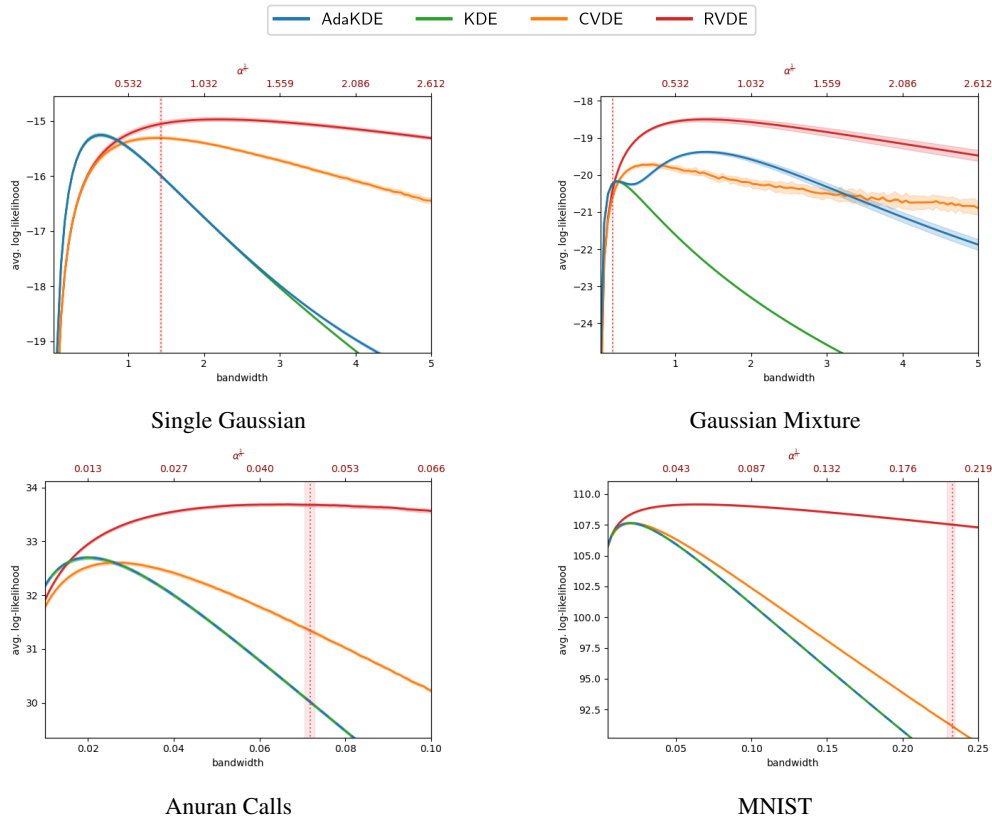|	RVDE	CVDE	KDE	AdaKDE
Gaussian	0.0376	0.265	0.0340	0.266
Anuran Calls	0.0581	0.490	0.0787	0.870
MNIST	17.4	408	12.5	75.0

Figure 6: Comparison of the estimators as the bandwidth varies. All the estimators implement the rational kernel.

The plot in Figure 6 displays the (test) log-likelihood for all the estimators as the bandwidth hyperparameter h varies (see the definition of the baselines in Section 5.1). In order to compare RVDE on the same scale as the other estimators, we convert h to α via:

$$\alpha = \int_0^\infty K\left(\frac{t}{h}\right) \, \mathrm{d}t = h^n \int_0^\infty K(t) \, \mathrm{d}t. \tag{18}$$

As can be seen, RVDE outperforms the baselines (each with its respective best bandwidth) in all the cases considered. The margin between RVDE and the baselines is especially evident on the more complex and high-dimensional datasets (Anuran Calls and MNIST). This confirms that the geometric benefits and the continuity properties of RVDE translate into better estimates for densities of different nature and increasing dimensionality.

Table 1 reports the average runtime for an experimental run (with a single fixed bandwidth) for each estimator. RVDE outperforms the CVDE as well as AdaKDE by an extremely large margin. KDE achieves comparable runtimes to RVDE: it is slightly faster on Gaussian and MNIST while it is slightly slower on Anuran Calls. This confirms empirically the discussion from Section 4.2: RVDE is signifi-

cantly more efficient than CVDE and has the same (asymptotic) complexity as KDE.

Table 2: Standard deviations of the (test) log-likelihood over 5 experimental runs. RVDE is highlighted in blue. Each estimator is considered with its best bandwidth.

	RVDE	CVDE	KDE	AdaKDE
Gaussian	0.788	0.843	0.572	0.572
Anuran Calls	1.170	1.253	1.152	1.152
MNIST	5.507	5.767	5.735	5.735

Table 2 separately reports the standard deviation of the log-likelihood (averaged over P_{test}) w.r.t. the dataset sampling. For each estimator, we consider its best bandwidth according to the results from Figure 6. We first observe that RVDE achieves lower standard deviation than CVDE on all the datasets. This corroborates the hypothesis that the continuity of RVDE results in more stable estimates than those obtained by the highly-discontinuous CVDE. KDE and AdaKDE achieve the lowest standard deviations on the

Gaussian and Anuran Calls datasets. This is likely due to the smoothness of such estimators and again confirms the benefit of regularity biases in terms of stability. However, on the most complex dataset considered (MNIST) RVDE outperforms the baselines. This suggests that for articulated densities the biases of geometric nature become more beneficial than generic biases such as smoothness.

6 CONCLUSIONS AND FUTURE WORK

In this work we introduced a non-parametric density estimator (RVDE) benefiting from the geometric properties of Voronoi tessellations while being continuous and computationally efficient. We provided both theoretical and empirical investigations of RVDE.

An interesting line for future investigation is to explore the radial construction of RVDE on Riemannian manifolds beyond the Euclidean space. In this generality the rays correspond to geodesics defined via the exponential map of the given manifold. A variety of Riemannian manifolds arise in statistics and machine learning. For example, data on spheres are the object of study of directional statistics (Mardia et al. 2009), hyperbolic spaces are routinely deployed to represent hierarchical data (Nickel et al. 2017) and complex projective spaces correspond to Kendall shape spaces from computer vision (Klingenberg 2020). Those areas of research can potentially benefit from the geometric characteristics and the computational efficiency of an extension of RVDE to Riemannian manifolds.

7 Acknowledgements

References

Baxter, Michael J., Christian C. Beardah, and Richard V. S. Wright (1997). "Some archaeological applications of kernel density estimates". In: *Journal of Archaeological Science* 24.4, pp. 347–354.

Bo, Hu et al. (2017). "Wind speed model based on kernel density estimation and its application in reliability assessment of generating systems". In: *Journal of Modern Power Systems and Clean Energy* 5.2, pp. 220–227.

Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Carr, Daniel B., Anthony R. Olsen, and Denis White (1992). "Hexagon mosaic maps for display of univariate and bivariate geographical data". In: *Cartography and Geographic Information Systems* 19.4, pp. 228–236.

Comaniciu, Dorin and Peter Meer (2002). "Mean shift: A robust approach toward feature space analysis". In: *IEEE Transactions on pattern analysis and machine intelligence* 24.5, pp. 603–619.

Corless, Robert M. et al. (1996). "On the LambertW function". In: *Advances in Computational mathematics* 5.1, pp. 329–359.

Dehnad, Khosrow (1987). *Density estimation for statistics and data analysis*.

Deng, Li (2012). "The mnist database of handwritten digit images for machine learning research". In: *IEEE Signal Processing Magazine* 29.6, pp. 141–142.

Devroye, L. P. and T. J. Wagner (1979). "The L1 convergence of kernel density estimates". In: *The Annals of Statistics*, pp. 1136–1139.

Diggle, Peter J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.

Dua, Dheeru and Casey Graff (2017). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml.

Duyckaerts, Charles, Gilles Godefroy, and Jean-Jacques Hauw (1994). "Evaluation of neuronal numerical density by Dirichlet tessellation". In: *Journal of neuroscience methods* 51.1, pp. 47–69.

Ebeling, H. and G. Wiedenmann (1993). "Detecting structure in two dimensions combining Voronoi tessellation and percolation". In: *Physical Review E* 47.1, p. 704.

Freedman, David and Persi Diaconis (1981). "On the histogram as a density estimator: L 2 theory". In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57.4, pp. 453–476.

Gabriel, K. Ruben and Robert R. Sokal (1969). "A new statistical approach to geographic variation analysis". In: *Systematic zoology* 18.3, pp. 259–278.

Gramacki, Artur (2018). *Nonparametric kernel density estimation and its computational aspects*. Springer.

Izenman, Alan Julian (1991). "Review papers: Recent developments in nonparametric density estimation". In: *Journal of the american statistical association* 86.413, pp. 205–224.

Klingenberg, Christian Peter (2020). "Walking on Kendall's shape space: understanding shape spaces and their coordinate systems". In: *Evolutionary Biology* 47.4, pp. 334–352.

Lee, Jasper C. H. et al. (2021). "Finding an approximate mode of a kernel density estimate". In: *29th Annual European Symposium on Algorithms (ESA 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Mardia, Kanti V. and Peter E. Jupp (2009). *Directional statistics*. Vol. 494. John Wiley & Sons.

Marron, J. S. (1987). "A comparison of cross-validation techniques in density estimation". In: *The Annals of Statistics*, pp. 152–162.

Masry, Elias (1997). "Multivariate probability density estimation by wavelet methods: Strong consistency and

rates for stationary time series". In: *Stochastic processes and their applications* 67.2, pp. 177–193.

Nickel, Maximillian and Douwe Kiela (2017). "Poincaré embeddings for learning hierarchical representations". In: *Advances in neural information processing systems* 30.

Okabe, Atsuyuki et al. (2009). *Spatial tessellations: concepts and applications of Voronoi diagrams*. Vol. 501. John Wiley & Sons.

Ord, J. K. (1978). "How many trees in a forest". In: *Mathematical Scientist* 3, pp. 23–33.

Parzen, Emanuel (1962). "On estimation of a probability density function and mode". In: *The annals of mathematical statistics* 33.3, pp. 1065–1076.

Pearson, Karl (1894). "Contributions to the mathematical theory of evolution". In: *Philosophical Transactions of the Royal Society of London. A* 185, pp. 71–110.

Polianskii, Vladislav et al. (2022). "Voronoi density estimator for high-dimensional data: Computation, compactification and convergence". In: *Uncertainty in Artificial Intelligence*. PMLR, pp. 1644–1653.

Rosenblatt, Murray (1956). "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27.3, pp. 832–837. ISSN: 00034851. URL: http : / / www . jstor . org / stable/2237390.

Scott, David W. (1988). "A note on choice of bivariate histogram bin shape". In: *Journal of Official Statistics* 4.1, p. 47.

– (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Vannucci, Marina (1995). *Nonparametric density estimation using wavelets*. Institute of Statistics & Decision Sciences, Duke University.

Vavilova, Iryna et al. (2021). "The Voronoi tessellation method in astronomy". In: *Intelligent Astrophysics*. Springer, pp. 57–79.

Walt, Christiaan Maarten van der and Etienne Barnard (2017). "Variable kernel density estimation in high-dimensional feature spaces". In: *Thirty-first AAAI conference on artificial intelligence*.

Walter, G. G. (1995). "Estimation with wavelets and the curse of dimensionality". In: *Manuscript, Department of Mathematical Sciences, University of Wisconsin-Milwaukee*.

Wand, Matt P., M. Chris Jones, et al. (1994). "Multivariate plug-in bandwidth selection". In: *Computational Statistics* 9.2, pp. 97–116.

Wang, Bin and Xiaofeng Wang (2007). "Bandwidth selection for weighted kernel density estimation". In: *arXiv preprint arXiv:0709.1616*.

Xie, Zhixiao and Jun Yan (2008). "Kernel density estimation of traffic accidents in a network space". In: *Computers, environment and urban systems* 32.5, pp. 396–406.

An Efficient and Continuous Voronoi Density Estimator: Supplementary Materials

PROOFS OF RESULTS FROM SECTION 4.3

Proposition 7.1. *Fix $l > 0$ and suppose $K \in C^1(\mathbb{R}_{>A})$. Then the iteration β_{m+1} of the Newton-Raphson method for computing $\beta(l)$ by solving Equation 7 takes form:*

$$\beta_{m+1} = \beta_m \left(1 + \frac{1}{n} \left(1 - \frac{l^n K(\beta_m l) - n\alpha}{l^n K(\beta_m l) - n \int_0^l t^{n-1} K(\beta_m t) \, \mathrm{d}t} \right) \right). \tag{19}$$

Moreover, if K is convex then the Newton-Raphson method converges for any initial value β_0 i.e., $\lim_{m \to \infty} \beta_m = \beta(l)$.

Proof. Consider

$$F(\beta) = \int_0^l t^{n-1} K(\beta t) \, \mathrm{d}t - \alpha. \tag{20}$$

The iteration of the Newton-Rhapson method for solving $F(\beta) = 0$ takes form:

$$\beta_{m+1} = \beta_m - \frac{F(\beta_m)}{\frac{\mathrm{d}F}{\mathrm{d}\beta}(\beta_m)}. \tag{21}$$

Via integration by parts we obtain:

$$\frac{\mathrm{d}F}{\mathrm{d}\beta}(\beta) = \int_0^l t^n \frac{\mathrm{d}K}{\mathrm{d}t}(\beta t) \, \mathrm{d}t = \frac{1}{\beta} \left(l^n K(\beta l) - n \int_0^l t^{n-1} K(\beta t) \, \mathrm{d}t \right). \tag{22}$$

Equation 19 follows then from Equation 21 by elementary algebraic manipulations. The convergence guarantee follows from the fact that if K is convex then F is easily seen to be convex as well. The Newton-Raphson method is well-known to be convergent for convex functions (Boyd et al. 2004). $\square$

Proposition 7.2. *The function $\beta : \mathbb{R}_{>0} \to \mathbb{R}$ is increasing, has a zero at $l = (n\alpha)^{\frac{1}{n}}$ and has an horizontal asymptote:*

$$\lim_{l \to +\infty} \beta(l) = \left(\frac{1}{\alpha} \int_0^\infty t^{n-1} K(t) \, \mathrm{d}t \right)^{\frac{1}{n}}. \tag{23}$$

Moreover if $K \in C^1(\mathbb{R}_{>A})$ then $\beta \in C^1(\mathbb{R}_{>0})$ and it satisfies the differential equation:

$$\left(l - \frac{n\alpha}{l^{n-1} K(\beta(l)l)} \right) \frac{\mathrm{d}\beta}{\mathrm{d}l}(l) = -\beta(l). \tag{24}$$

Proof. The claim on the monotonicity of β follows directly from its definition (Equation 7) and the hypothesis that K is decreasing. In order to compute its zero, note that $\beta(l) = 0$ implies $\alpha = \int_0^l K(0) t^{n-1} = \frac{l^n}{n}$ and thus $l = (n\alpha)^{\frac{1}{n}}$. For the asymptote note that for $l = +\infty$ Equation 7 becomes by a change of variables:

$$\int_0^\infty t^{n-1} K(\beta(+\infty)t) \, \mathrm{d}t = \frac{1}{\beta(+\infty)^n} \int_0^\infty t^{n-1} K(t) \, \mathrm{d}t = \alpha. \tag{25}$$

Lastly, in order to obtain the differential equation for β we differentiate Equation 7 on both sides and get:

$$0 = \frac{\mathrm{d}}{\mathrm{d}l} \int_0^l t^{n-1} K(\beta(l)t) \, \mathrm{d}t = l^{n-1} K(\beta(l)l) + \int_0^l t^{n-1} \frac{\mathrm{d}}{\mathrm{d}l} K(\beta(l)t) \, \mathrm{d}t$$

$$= l^{n-1} K(\beta(l)l) + \frac{\mathrm{d}\beta}{\mathrm{d}l}(l) \int_0^l t^n \frac{\mathrm{d}K}{\mathrm{d}t}(\beta(l)t) \, \mathrm{d}t \tag{26}$$

$$= l^{n-1} K(\beta(l)l) + \frac{\mathrm{d}\beta}{\mathrm{d}l}(l) \frac{l^n K(\beta(l)l) - n\alpha}{\beta(l)}$$

where in the first identity we deployed the (distributional) Leibniz rule while in the last one we deployed integration by parts.

$\square$

Proposition 7.3. *The modes of f_P are as follows:*

(1) $p \in P$ if $d(p,q) > 2\varepsilon$ for every Voronoi cell $C(q)$ adjacent to $C(p)$,

(2) $\frac{p+q}{2}$ for $p,q \in P$ if $\frac{p+q}{2} \in C(p) \cap C(q)$ and $d(p,q) < 2\varepsilon$,

(3) all points belonging to the segment $[p,q]$ for $p,q \in P$ if $\frac{p+q}{2} \in C(p) \cap C(q)$ and $d(p,q) = 2\varepsilon$.

Proof. Pick $p \in P$. If p satisfies the hypothesis of the first claim then $l(x) > \varepsilon$ for every $x \in C(p)$ and thus $\beta(l(x)) > 0$ by Proposition 7.2. Since K is decreasing, f_P decreases radially w.r.t. p in $C(p)$ and the first claim follows. If p does not satisfy the hypothesis of the first claim then $\beta(l(x)) \leq 0$ for some $x \in C(p)$. With the exception of the case $\beta(l) = 0$, the modes lie then on the boundary and are of the form $K(\beta(l)l)$ up to a multiplicative constant. The function $\beta(l)l$ is increasing in l since by appealing to Proposition 7.2 we can compute its derivative:

$$\frac{\mathrm{d}\beta(l)l}{\mathrm{d}l} = \beta(l) + l\frac{\mathrm{d}\beta(l)}{\mathrm{d}l} = \beta(l)\frac{n\alpha}{n\alpha - l^n K(\beta(l)l)} \geq 0. \tag{27}$$

Since $l(x)$ has local minima at midpoints of segments connecting points in P, $K(\beta(l)l)$ is locally maximized therein and the second claim follows. In the hypothesis of the third claim β vanishes on the segment and the density is thus constant.

$\square$

ADDITIONAL EXPERIMENTS

In this section we report additional experimental results complementing the ones in the main of the manuscript. For completeness, we evaluate the density estimators on different kernels. Figure 7 displays comparative results for all the estimators with the exponential and the Gaussian kernel (note that the latter does not apply to RVDE). Moreover, we experiment with different dimensions and evaluation metrics other than average log-likelihood. This is possible only on a synthetic dataset where the dimension can vary and where the ground-truth density ρ is known. The latter is necessary for the metric considered. Figure 8 displays a comparison on a high-dimensional Gaussian mixture ($n = 30$) as well as a comparison on the Gaussian mixture as in Section 5 ($n = 10$) where the evaluation metric is the empirical *Hellinger distance* on the test set:

$$\frac{1}{2|P_{\text{test}}|} \sum_{x \in P_{\text{test}}} \left(f_P(x)^{\frac{1}{2}} - \rho(x)^{\frac{1}{2}} \right)^2. \tag{28}$$
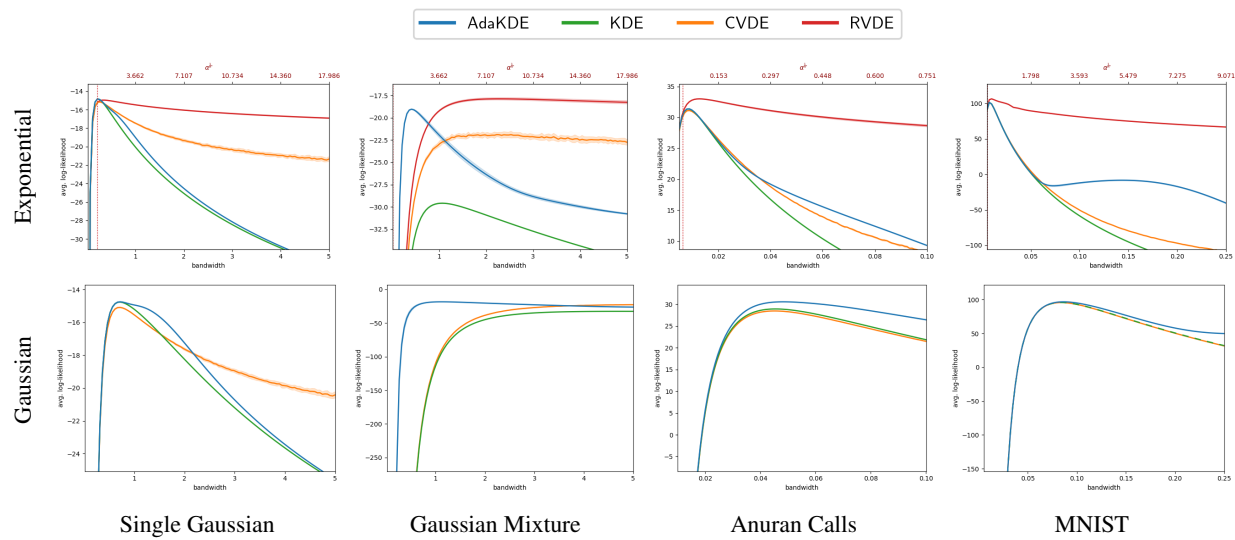
Figure 7: Comparison of the estimators with the exponential and Gaussian kernel as the bandwidth varies.
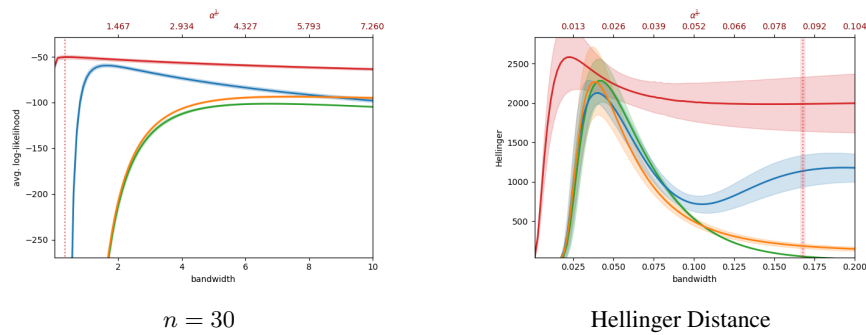

Figure 8: Comparison of the estimators on a 30-dimensional Gaussian mixture (left) and on a 10-dimensional Gaussian mixture with the Hellinger distance as a metric (right).