Winner-takes-all learners are geometry-aware conditional density estimators

Victor Letzelter [*1 2] **David Perera** [*2] **Cédric Rommel** [3]
Mathieu Fontaine [2] **Slim Essid** [2] **Gaël Richard** [2] **Patrick Pérez** [4]

Abstract

Winner-takes-all training is a simple learning paradigm, which handles ambiguous tasks by predicting a set of plausible hypotheses. Recently, a connection was established between Winner-takes-all training and centroidal Voronoi tessellations, showing that, once trained, hypotheses should quantize optimally the shape of the conditional distribution to predict. However, the best use of these hypotheses for uncertainty quantification is still an open question. In this work, we show how to leverage the appealing geometric properties of the Winner-takes-all learners for conditional density estimation, without modifying its original training scheme. We theoretically establish the advantages of our novel estimator both in terms of quantization and density estimation, and we demonstrate its competitiveness on synthetic and real-world datasets, including audio data.

1. Introduction

Machine-learning-based predictive systems are faced with a fundamental limitation when there is some ambiguity in the data or in the task itself. This results in a non-deterministic relationship between inputs and outputs, which is challenging to cope with. Characterizing this inherent uncertainty is the problem of conditional distribution estimation.

The recently introduced Winner-takes-all (WTA) training scheme (Guzman-Rivera et al., 2012; Lee et al., 2016) is a novel approach addressing ambiguity in machine learning. This scheme leverages several models, generally a neural network equipped with several heads, to produce multiple predictions, also called *hypotheses*. It trains these hypotheses competitively, updating only the hypothesis that yields the current best prediction. Experimental evidence has demonstrated that this approach enhances the diversity of predictions, with each head gradually specializing in a subset of the data distribution.

At the same time, a limited body of work has tried to theoretically elucidate the appealing characteristics of Winner-takes-all learners. Specifically, Rupprecht et al. (2017) described the geometrical properties of the trained WTA learners using the formalism of *centroidal Voronoi tessellations*. This approach is linked to the field of quantization, where the objective is to represent an arbitrary distribution optimally using a finite set of points (Zador, 1982).

Being able to quantize a distribution in an input-dependent manner, WTA learners have the potential to model the *geometric* information of a distribution. This raises the following question: can WTA learners be used to make accurate *probabilistic* predictions? This paper affirms this possibility.

We build upon the recent findings of Letzelter et al. (2023), which proposed modeling uncertainty from WTA predictions, using either Dirac or uniform mixtures. We extend this idea by proposing a kernel-based density estimator for WTA predictors. This enables the computation of uncertainty metrics, such as the negative log-likelihood, from trained WTA models. This development introduces a novel method for the probabilistic evaluation of WTA predictions. Notably, it can be used even when only a single target from the conditional distribution is available for each input.

The key contributions of this work are as follows:

1. We introduce an estimator that provides a comprehensive probabilistic interpretation of WTA predictions while retaining their appealing geometric properties.
2. We mathematically validate the competitiveness of our estimator, both in terms of geometric quantization properties and probabilistic convergence, as the number of hypotheses increases.
3. We empirically substantiate our estimator through experiments on both synthetic and real-world data, including audio signals.[1]

[*]Equal contribution. [1]Valeo.ai, Paris, France [2]LTCI, Télécom Paris, Institut Polytechnique de Paris, France [3]Meta AI, Paris, France [4]Kyutai, Paris, France. Correspondence to: Victor Letzelter <victor.letzelter@telecom-paris.fr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Code at https://github.com/Victorletzelter/VoronoiWTA.

2. Background

2.1. Winner-takes-all training

The Winner-takes-all training scheme is the basic building block of the *Multiple Choice Learning* family of approaches (Guzman-Rivera et al., 2012; Lee et al., 2016; 2017; Tian et al., 2019). It was introduced to deal with inherently ambiguous prediction tasks. Specifically, we are not only interested in predicting a single output $f_\theta(x) \in \mathcal{Y}$ from a given input $x \in \mathcal{X}$ (f_θ can typically be a deep neural network with parameters θ). Instead, we want to perform several predictions $f_\theta^1(x), \ldots, f_\theta^K(x)$ accounting for K potential outcomes.

More precisely, let $f_\theta \triangleq (f_\theta^1, \ldots, f_\theta^K) \in \mathcal{F}(\mathcal{X}, \mathcal{Y}^K)$, which could be for instance a multi-head deep neural network, and let $(x, y) \in \mathcal{X} \times \mathcal{Y}$ be a pair sampled from a joint distribution $\mathbb{P}$ (with density $\rho(x, y)$).

In a supervised setting, the WTA training consists in:

1. performing a forward pass through the loss ℓ for all predictors,
2. then backpropagating the loss gradients for the selected *winner* hypothesis:

$$\mathcal{L}^{\mathrm{WTA}}(\theta) \triangleq \min_{k \in [\![1,K]\!]} \ell\left(f_\theta^k(x), y\right) . \tag{1}$$

This two-step approach, originally proposed by Lee et al. (2016), makes it possible to use gradient-based optimization, despite the non-differentiability of the $\min$ operator in (1).

2.2. Desirable geometrical properties

Rupprecht et al. (2017) have shown that, once trained, the output of the set of predictors $(f_\theta^1(x), \cdots, f_\theta^K(x))$ can be interpreted as an *input-dependent centroidal Voronoi tessellation*, thereby providing a geometrical probabilistic interpretation of WTA. As done by Rupprecht et al. (2017), we study the case where $\ell(\hat{y}, y) = \|\hat{y} - y\|^2$ is the the L^2 loss. In a standard machine learning setting, where a single prediction is provided, one can prove that the risk

$$\mathbb{E}_{(x,y)\sim\rho(x,y)}[\ell(f_\theta(x), y)] , \tag{2}$$

is minimized when $\forall x \in \mathcal{X},\ f_\theta(x) = \mathbb{E}[Y_x]$, noting $Y_x \sim \mathbb{P}_x$ the conditional distribution and ρ_x its density. The proof of this result is based on the customary assumption that the predictor f_θ is sufficiently expressive, so that minimizing the risk (2) is equivalent to minimizing the input-dependent risk, $\mathbb{E}_{y\sim\rho_x(y)}[\ell(f_\theta(x), y)]$, for each fixed input x. When multiple predictors are used, as in the WTA case, the situation is more complex. In this case, after defining Voronoi cells as:

$$\mathcal{Y}_k(g) \triangleq \{y \in \mathcal{Y} \mid \ell(g_k, y) < \ell(g_r, y), \forall r \neq k\} , \tag{3}$$

for some arbitrary set of generators $(g_1, \cdots, g_K)$, the input-dependent risk writes, for each $x \in \mathcal{X}$, as

$$\sum_{k=1}^{K} \int_{\mathcal{Y}_\theta^k(x)} \ell(f_\theta^k(x), y)\rho_x(y)\mathrm{d}y , \tag{4}$$

where for ease of notation $\mathcal{Y}_\theta^k(x) \triangleq \mathcal{Y}_k(f_\theta(x))$. Note that (4) is not differentiable with respect to the parameters θ, which are involved both in the integrand and in the integration domain. This issue can be alleviated by uncoupling the two variables and defining:

$$\mathcal{K}(g, z) \triangleq \sum_{k=1}^{K} \int_{\mathcal{Y}_k(g)} \ell(z_k, y)\, \rho_x(y)\mathrm{d}y , \tag{5}$$

where $z = (z_1, \cdots, z_K)$. Note that (4) corresponds to $\mathcal{K}(f_\theta(x), f_\theta(x))$.

For the purpose of the next proposition, let us define a centroidal Voronoi tessellation.

Definition (Centroidal Voronoi Tessellation)**.** *We say that $\{\mathcal{Y}_k(z)\}$ forms a centroidal Voronoi tessellation with respect to a density function ρ_x and a loss function ℓ if, for each cell k, the generator z_k minimizes the weighted loss over its region:*

$$\int_{\mathcal{Y}_k(z)} \rho_x(y)\ell(z_k, y)\,\mathrm{d}y = \inf_{z' \in \mathcal{Y}_k^\star(z)} \int_{\mathcal{Y}_k(z)} \rho_x(y)\ell(z', y)\mathrm{d}y ,$$

where $\mathcal{Y}_k^\star(z)$ is the closure of $\mathcal{Y}_k(z)$.

Based on formulation (5), Rupprecht et al. (2017) show that one can adapt the following result.

Proposition 2.1 (Du et al., 1999)**.** *A necessary condition for minimizing (5) is that $\mathcal{Y}_k(g)$ are the Voronoi regions generated by the z_k, and simultaneously, $\{\mathcal{Y}_k(z)\}$ forms a centroidal Voronoi tessellation generated by $\{z_k\}$.*

In particular, if ℓ is the L^2-loss, this condition implies that, for each non-zero probability cell, the optimal hypotheses placements correspond to cell-restricted conditional expectations as stated in Theorem 1 of Rupprecht et al. (2017):

$$f_\theta^k(x) = \mathbb{E}[Y_x \mid Y_x \in \mathcal{Y}_\theta^k(x)] . \tag{6}$$

This necessary condition, which offers a geometrical interpretation of the Winner-takes-all optimum, has been verified experimentally in previous works (Rupprecht et al., 2017; Letzelter et al., 2023), thus demonstrating the method's potential to predict *input-dependent* centroidal Voronoi tessellations using deep neural networks.

2.3. Probabilistic interpretation as a mixture model

Proposition 2.1 highlights the geometric advantages of WTA, but it does not provide a full probabilistic interpretation of this method. First, (6) is valid only in Voronoi

cells with strictly positive mass, *i.e.*, containing at least one sample from the ground-truth empirical distribution. Furthermore, the WTA predictor from Rupprecht et al. (2017) implicitly affects equal probability to all Voronoi cells, regardless of their ground-truth probability mass.

As a possible solution, Tian et al. (2019) and Letzelter et al. (2023) propose to additionally train *score* heads $\gamma_\theta^1, \dots, \gamma_\theta^K \in \mathcal{F}(\mathcal{X}, [0,1])$, estimating the probability mass of each cell $\mathbb{P}(Y_x \in \mathcal{Y}_\theta^k(x))$ by jointly optimizing in θ the WTA loss (1) with the cross-entropy

$$\mathcal{L}^{\text{scoring}}(\theta) \triangleq \sum_{k=1}^{K} \text{BCE}\left(\mathbb{1}\left[y \in \mathcal{Y}_\theta^k(x)\right], \gamma_\theta^k(x)\right), \quad (7)$$

between the predicted assignation probability $\gamma_\theta^k(x)$ and the actual assignation, where $\text{BCE}(p,q) \triangleq -p\log(q) - (1-p)\log(1-q)$. The full training objective is therefore defined as a compound loss $\mathcal{L}^{\text{WTA}} + \beta\mathcal{L}^{\text{scoring}}$. Mirroring Proposition 2.1, one can show that a necessary condition to minimize the scoring objective is that each $\gamma_\theta^k(x)$ is an unbiased estimator of the probability mass of its cell:

$$\gamma_\theta^k(x) = \mathbb{P}(Y_x \in \mathcal{Y}_\theta^k(x)). \quad (8)$$

Assuming now that (6) and (8) are verified after training, through the minimization of the combined objective, Letzelter et al. (2023) argued that it is possible to interpret the outputs of such a model probabilistically, as a Dirac mixture:

$$\hat{\rho}_x(y) = \sum_{k=1}^{K} \gamma_\theta^k(x)\delta_{f_\theta^k(x)}(y). \quad (9)$$

Let $\hat{Y}_x \sim \hat{\rho}_x$ denote the random variable sampled from this estimated conditional distribution. The Dirac mixture interpretation (9) has at least two desired properties:

1. **[centroidal property]** the cell-restricted expectation with respect to the estimated distribution matches the ground truth:

$$\mathbb{E}[\hat{Y}_x \mid \hat{Y}_x \in \mathcal{Y}_\theta^k(x)] = \mathbb{E}[Y_x \mid Y_x \in \mathcal{Y}_\theta^k(x)], \quad (10)$$

2. **[cell-scoring property]** the predicted probability mass of the Voronoi cells is unbiased:

$$\mathbb{P}(\hat{Y}_x \in \mathcal{Y}_\theta^k(x)) = \mathbb{P}(Y_x \in \mathcal{Y}_\theta^k(x)). \quad (11)$$

This interpretation is hence appealing as it captures the global shape of the distribution. However, it presents a major caveat: it does not capture local variations of mass within the Voronoi cells.

3. Limitations of current estimators

We illustrate hereafter the main limitations of the probabilistic interpretation of the score-based WTA proposed in Letzelter et al. (2023). To this end, we consider a toy example inspired from Rupprecht et al. (2017).

Our goal is to predict an input-dependent distribution $\hat{\rho}_x(y)$, where x lives in the unit-segment $\mathcal{X} = [0,1]$, and y is restricted to the 2D-square $\mathcal{Y} = [-1,1]^2$. The latter is split into four quadrants: $S_1 = [-1,0) \times [-1,0), S_2 = [-1,0)\times[0,1], S_3 = [0,1]\times[-1,0)$ and $S_4 = [0,1]\times[0,1]$. The target distribution for each x is then generated by first sampling one of the four quadrants with probabilities $p(S_1) = p(S_4) = \frac{1-x}{2}$ and $p(S_2) = p(S_3) = \frac{x}{2}$. Once a region is sampled, a point is then drawn from a predetermined distribution restricted to that region: uniform distributions in S_1 and S_4, and Gaussian distributions in S_2 and S_3 (with different standard deviations, respectively, $\sigma_2 \gg \sigma_3$).

We trained a 20-hypothesis scoring-based WTA model, consisting of a three-layer MLP, on this dataset. The predictions for three inputs $x \in \{0.01, 0.6, 0.9\}$ are shown in Figure 1.

For small values of x, the ground-truth conditional distribution is piece-wise uniform. In this situation, the mass does not vary much inside the predicted Voronoi cells, and the whole distribution is hence well summarized by the predicted hypotheses and scores alone. However, the same cannot be said as x increases. Indeed, when $x \in \{0.6, 0.9\}$, we see on the bottom-right quadrant (S_3) that the small-variance Gaussian is modeled by a single hypothesis and Voronoi cell. Although the hypothesis seems to be well-positioned at the true Gaussian mean and its corresponding cell seems to verify both centroidal and cell-scoring properties, the local mass variations within the cell are not well-described. More precisely, neither a Dirac delta, as in (9), nor a cell-restricted uniform distribution seem like good estimations of the underlying conditional probability density within S_3.

This problem of intra-cell density approximation can be partially mitigated by increasing the number of hypotheses and using uniform mixtures. However, we argue that more accurate estimators can be built from the adaptive grid provided by WTA, even when the number of hypotheses is low.

This example highlights the need for WTA-based conditional density estimators taking into account the data distribution geometry through optimal hypotheses placement.

4. Conditional density approximation

The goal of this work is to propose a probabilistic interpretation of the Winner-takes-all predictions as a conditional density estimator that preserves the global geometric prop-

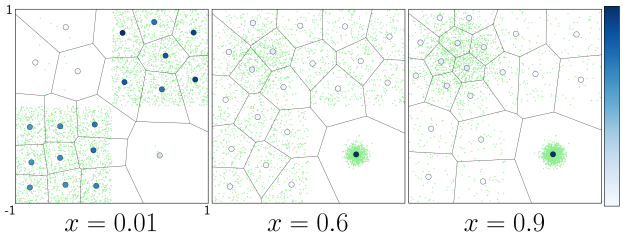

Figure 1. **Limitations of Dirac Mixtures.** Model predictions for different inputs x (columns) are shown with blue-shaded circles; the colorbar indicates hypothesis scores. Green points depict the target distribution for each input. Black lines mark the boundaries of the Voronoi tessellation associated with the predictions.

erties of the predictions (centroidal Voronoi tessellation) and captures the local variations of the probability density, including inside Voronoi cells. Ideally, we would also like our estimators to verify both centroidal (10) and cell-scoring properties (11).

4.1. Kernel WTA

A straightforward way to model intra-cell density variations is to place a kernel $K_h(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ on each hypothesis k as in a traditional Parzen estimator (Rosenblatt, 1956; Parzen, 1962). This procedure defines the following conditional density estimator, called *Kernel-WTA* hereafter:

$$\hat{\rho}_x(y) = \sum_{k=1}^{K} \gamma_\theta^k(x) K_h(f_\theta^k(x), y)\,, \tag{12}$$

where $h \in \mathbb{R}_+^*$ is the scaling factor of the kernel. In this work, we consider only isotropic kernels, *e.g.,* Gaussian, assumed to integrate to 1 in their second variable. Despite being simple and allowing intra-cell variations to be modeled, this method has drawbacks. Indeed, we see from (12) that, whenever h is too large, the kernels begin to diffuse density out of their Voronoi cells. As a result, *neither* the centroidal (10) nor the cell-scoring (11) properties hold anymore, meaning that the geometric advantages offered by the Winner-takes-all predictions are not fully preserved. As a second drawback, the convergence of Kernel-WTA is highly dependent on the choice of h, as discussed in Section 5.

4.2. Voronoi WTA

As mentioned in the previous section, the straightforward Kernel-WTA fails to preserve the geometric properties when h is too large. Inspired by Polianskii et al. (2022), we propose to alleviate this problem using truncated kernels. More precisely, let $V(g_k, K_h) \triangleq \int_{\mathcal{Y}_k(g)} K_h\left(g_k, \tilde{y}\right) \mathrm{d}\tilde{y}$ be the volume of the Voronoi cell defined by generator g_k, under the metric induced by kernel K_h. We define the *Voronoi-WTA* estimator as:

$$\hat{\rho}_x(y) = \sum_{k=1}^{K} \gamma_\theta^k(x) \frac{K_h\left(f_\theta^k(x), y\right)}{V(f_\theta^k(x), K_h)} \mathbb{1}\left(y \in \mathcal{Y}_\theta^k(x)\right)\,. \tag{13}$$

Unlike Kernel-WTA, the density estimations derived from (13) are designed to fulfill the cell-scoring property, *i.e.*, if $\hat{Y}_x \sim \hat{\rho}_x$, $\mathbb{P}(\hat{Y}_x \in \mathcal{Y}_\theta^k(x)) = \gamma_\theta^k(x) = \mathbb{P}(Y_x \in \mathcal{Y}_\theta^k(x))$. Using this property, the convergence in distribution of Voronoi-WTA as K approaches infinity, shown in Section 5.1, is independent of the choice of h. As in (12), h remains constant by design for each input x, and in each cell k. Note that increasing h causes a slight shift in the barycenter of the predicted distribution from the ground-truth expectation in each cell. Nonetheless, in our setup, h will be optimized after the optimization of θ, ensuring that (6) is still verified.

4.3. Likelihood computation and sampling

In practice, the use of the Voronoi-WTA defined in (13) raises three main questions: (1) how to sample from this estimator, (2) how to compute likelihoods, and (3) how to choose the scaling factor h.

Sampling. *Rejection sampling* is a simple way of sampling from (13). In practice, one can first draw a Voronoi cell $k \in [\![1, K]\!]$ from the discrete distribution of predicted scores $\{\gamma_\theta^l(x)\}$, then sample from the kernel $K_h(f_\theta^k(x), \cdot)$ until a sample falls in cell k. This approach was efficient enough for all our experiments. Note that whenever the number of hypotheses or the dimension is large, the more efficient *hit-and-run* sampling method from Polianskii et al. (2022) may be adapted for this setup.

Likelihood computation. The difficulty in computing (13) comes from the normalization term $V(f_\theta^k(x), K_h)$. In practice, it can be computed efficiently by re-writing $V(f_\theta^k(x), K_h)$ as a double-integral in spherical coordinates as explained in Polianskii et al. (2022). As the inner integral often has a closed-form solution for usual kernels, a simple Monte Carlo approximation of the outer integral allows us to efficiently estimate $V(f_\theta^k(x), K_h)$.

Choosing the scaling factor. We limit ourselves in this work to the case where the kernels in all Voronoi cells share the same scaling factor h. The latter has to be tuned, which can be done in practice based on the likelihood obtained by the model in a validation set. Note that when $h \to 0$, both Kernel-WTA (12) and Voronoi-WTA estimators (13) are equivalent to the Dirac mixture (9) from Letzelter et al. (2023). However, when h increases, Kernel-WTA loses the geometry captured by the hypotheses $\{f_\theta^k(x)\}$, while Voronoi-WTA preserves it, converging to a piecewise uniform distribution defined on the Voronoi tessellation. This is verified experimentally in Section 6.

5. Theoretical properties

In this section, we present informally the two main theoretical results of this work. These propositions are made precise in the appendix, together with other complementary results and their corresponding proofs.

5.1. Convergence in distribution independent of h

As a first theoretical contribution, we show in the following proposition that Voronoi-WTA is an effective conditional density estimator, in the sense that it converges to the ground-truth underlying distribution:

Proposition 5.1. *Under mild assumptions on $\mathcal{Y}$ and the data distribution, Voronoi-WTA (seen as a density estimator) converges in probability towards the conditional distribution $\mathbb{P}_x$ when the number K of hypotheses grows to infinity.*

A formal version of this result can be found in the appendix (Theorem B.10).

This result is similar to Theorem 4.1 by Polianskii et al. (2022) on the convergence of the Compactified Voronoi Density Estimator. Note, however, that our setting differs on two main points:

1. we study the more general problem of *conditional* densities,
2. our cell centroids correspond to the hypotheses predicted by a WTA estimator $z_k = f_\theta^k(x)$, while Polianskii et al. (2022) assume random generators $z_k \sim \rho_x$.

To deal with the last point, we assume that the underlying WTA estimator predicting $\{f_\theta^k(x), \gamma_\theta^k(x)\}$ has converged towards a global minimum of its WTA and scoring objectives. We give a sketch of the proof of this result below.

Proof. The convergence relies on a useful property of the Voronoi cells obtained when we minimize the WTA objective (1): their diameter vanishes as the number K of hypotheses grows to infinity. This observation is then used to prove the convergence.

Let $\mathbb{P}_K$ be the trained Voronoi-WTA estimator with K hypotheses. By the Portmanteau Lemma (Van der Vaart, 2000), it is sufficient to show that $\mathbb{P}_K(E) \to \mathbb{P}(E)$ as $K \to +\infty$ for any measurable set $E \subseteq \mathcal{Y}$ such that $\lambda(\partial E) = 0$, where λ denotes the Lebesgue measure. If we fix K, the Voronoi tiling $(\mathcal{Y}_\theta^k)_{k\in[\![1,K]\!]}$ induces a partition of E. Accordingly, we split E as the disjoint union $E = E^{\text{int}} \cup E^{\text{ext}}$, where E^{int} denotes the Voronoi cells included in E, and E^{ext} the Voronoi cells intersecting its border ∂E. Now, since the radius of each cell $\mathcal{Y}_\theta^k$ asymptotically vanishes (Proposition B.9), E^{ext} is concentrated on the border ∂E and is therefore negligible. The proof is concluded by observing that $\mathbb{P}_K$ and $\mathbb{P}$ coincide on E^{int} (cell-scoring property (11)).

Note that this last argument does not hold for Kernel-WTA, as the cell-scoring property does not hold for this model. Also, note that Proposition 5.1 requires no assumptions on the choice of the scaling factor h. This is another advantage of Voronoi-WTA, this time in terms of uncertainty modeling.

5.2. Better asymptotic quantization

Voronoi-WTA estimators model the conditional distribution using an adaptive grid. To measure how well a finite set of points $\mathcal{Z} = \{z_k\}_{k\in[\![1,K]\!]}$ approximates a data distribution $\mathbb{P}_x$, it is customary to use the *quantization error*, also called quadratic risk or quadratic distortion (Pagès & Printems, 2003):

$$\mathcal{R}(\mathcal{Z}) = \int_{\mathcal{Y}} \min_{z\in\mathcal{Z}} \|y - z\|_2^2 \, \rho_x(y) \mathrm{d}y \,. \tag{14}$$

A natural baseline that we can use to evaluate the advantage of the adaptative grid provided by Voronoi-WTA is the regular grid, which we call *Histogram* hereafter (*e.g.,* Imani & White (2018)). Note that the regular grid is a particular case of a Voronoi tessellation.

The quantization error is notoriously hard to study in the general case (Graf & Luschgy, 2007). However, things become amenable to analysis in the asymptotic regime. With this in mind, Zador's theorem (Zador, 1982), a powerful result from quantization theory, can be used to describe the asymptotic optimal quantization error. We sum up our observations in the following statement (see Propositions B.12 and B.13 for a complete formulation).

Proposition 5.2. *Under mild regularity assumptions, denoting $d = \dim(\mathcal{Y})$, J_d a constant depending only on the dimension, $\mathrm{vol}(\mathcal{Y})$ the volume of $\mathcal{Y}$, and $\mathcal{Z}_x^{\mathrm{V}} = \{f_\theta^k(x)\}_{k\in[\![1,K]\!]}$, the quantization error has the following asymptotic equivalent as $K \to +\infty$:*

$$\mathcal{R}(\mathcal{Z}_x^{\mathrm{V}}) \sim J_d \left(\int_{\mathcal{Y}} \rho_x(y)^{\frac{d}{d+2}} \mathrm{d}y \right)^{\frac{d+2}{d}} \frac{1}{K^{2/d}} \,. \tag{15}$$

Denoting $\mathcal{Z}^{\mathrm{H}}$ the fixed grid points defining the Histogram baseline, we also have

$$\mathcal{R}(\mathcal{Z}^{\mathrm{H}}) \sim \frac{d}{12} \frac{\mathrm{vol}(\mathcal{Y})^{2/d}}{K^{2/d}} \,. \tag{16}$$

Note that, in the first order, the quantization error of the Histogram baseline only depends on the volume of the support of $\mathbb{P}_x$, whereas Voronoi-WTA takes into account local density information provided by ρ_x. This gives an insight into how the adaptative grid underlying Voronoi-WTA fits the geometry of the data distribution.

Furthermore, one can also observe that Voronoi-WTA and the Histogram have the same asymptotic rate of convergence, differing only by the leading constant. However, it can be

proved that this constant is smaller for Voronoi-WTA than for the Histogram baseline in most cases (Proposition B.14). Therefore, although the gap between $\mathcal{R}(\mathcal{Z}_x^{\mathrm{V}})$ and $\mathcal{R}(\mathcal{Z}^{\mathrm{H}})$ closes as the number K of hypotheses increases, Voronoi-WTA *always* has a strictly better quantization error, even asymptotically. This constitutes a real advantage of the adaptive grid provided by Voronoi-WTA over a static one and was empirically verified in Section 6.

6. Empirical study

The aim of this section is two-fold. First, it empirically justifies the relevance of the Voronoi-WTA-based conditional density estimators, compared to other possible designs based on WTA learners, such as Kernel-WTA. Second, it empirically validates the advantages of the Winner-takes-all training scheme against traditional baselines for conditional density estimation.

6.1. Experimental setting

We detail below our experimental settings. A more extensive description of design choices is deferred to Appendix C.

Datasets. We conducted experiments on four synthetic datasets, with $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = [-1, 1]^2$, as well as on the UCI benchmark (Hernández-Lobato & Adams, 2015).

- *Single Gaussian* corresponds to a single, isotropic, non-centered two-dimensional Gaussian which does not move as x varies.
- *Rotating Two Moons* is based on the two-moon dataset from SCIKIT-LEARN (Pedregosa et al., 2011), corresponding to entangled non-convex shapes. The target distribution was generated by rotating the latter with an angle $2\pi x$ for each $x \in [0, 1]$.
- *Changing Damier* is an adaptation of the dataset proposed in Rupprecht et al. (2017). It corresponds to a checkerboard of 16 squares, gradually interpolated towards its complementary checkerboard as x increases.
- *Uniform to Gaussians* is the illustrative dataset presented in Section 3.
- *UCI Regression datasets* (Dua & Graff, 2017) are a standard benchmark (Hernández-Lobato & Adams, 2015) to evaluate conditional density estimators.

WTA training framework. We used the WTA training scheme with scoring heads from Section 2.3. The density estimation was performed following the methodology described in Section 4, with uniform kernels and Gaussian kernels with several scales h.

Baselines. Two standard conditional density estimation baselines were considered in our experiments: Mixture Density Networks (MDN) (Bishop, 1994) and the Histogram (Imani & White, 2018) mentioned in Section 5.2. More details are given in Appendix C.1.1.

Architecture and training details. In each training setup with synthetic data, we used a three-layer MLP, with 256 hidden units. The Adam optimizer (Kingma & Ba, 2014) was used, and the models were trained until convergence of the training loss, using early stopping on the validation loss.

Metrics. To evaluate the performance of each model, we employed the Negative Log-Likelihood (NLL) and, when the target distribution is known, the Earth Mover's Distance (EMD). To assess how well each model preserved the geometry of the data distribution, we used the quantization risk, as defined in (14).

6.2. Qualitative analysis

Qualitative results are provided in Figure 2, where the predictions of Score-based WTA, Histogram, and MDN are compared. 'Score-based WTA' refers to both Voronoi-WTA and Kernel-WTA, which share the same predicted hypotheses and scores represented in the figure. Different behaviors can be observed for each of the three methods. For instance, in the Gaussian case, MDN predictions of mixture means collapse into a single point. This well-known *mode collapse* problem (Hjorth & Nabney, 1999; Graves, 2013; Rupprecht et al., 2017; Messaoud et al., 2018; Cui et al., 2019) is also observed for Rotated Two Moons and Changing Damier when the number of hypotheses increases. Concerning Histogram, we see on all datasets, except the Changing Damier, that it requires more hypotheses to reach the same resolution as the score-based WTA, which is able to optimally quantize the shape of all distributions with their predicted hypotheses.

6.3. Quantitative analysis on synthetic datasets

Our main quantitative results on the synthetic datasets are depicted in Figure 3.

Comparison to Histogram. One can see in Figure 3 that Histogram generally does not lead to competitive performance with respect to any metric unless a high number of hypotheses is used. An exception is observed in the case of the Changing Damier dataset where, by design, the Histogram aligns perfectly with the data when it is set to exactly 16 hypotheses (*cf.* Figure 2 right). When the number of hypotheses is large enough, the grid is sufficiently fine to represent the distribution geometry and Histogram's performance strongly improves. Note for the quantization error, however, that Histogram always performs worse than Voronoi-WTA, regardless of the number of hypotheses, as predicted by Proposition 5.2. These results showcase quantitatively the clear advantage of Voronoi-WTA's adaptive grid over Histogram.

Comparison to Mixture Density Networks. MDNs have

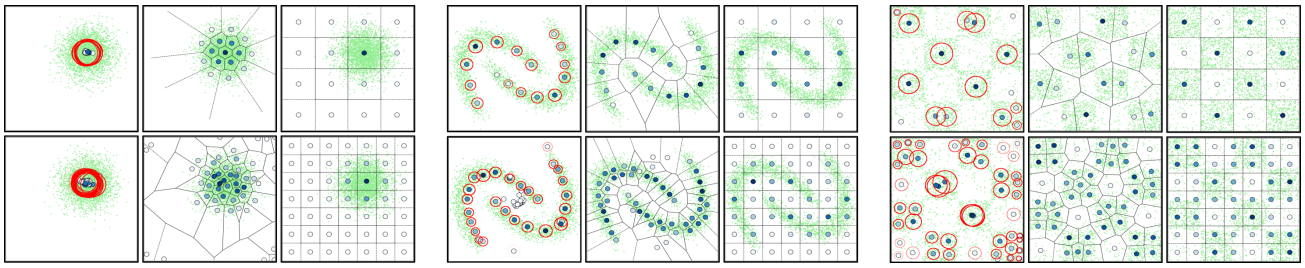

Figure 2. **Qualitative results.** Each panel shows a different dataset: Single Gaussian, Rotated Two Moons, and Changing Damier. Within each panel, columns correspond to predictions made by: MDN, Score-based WTA, and Histogram (left to right). Dots represent predicted (or fixed) hypotheses: means, centroids, and bins. Their colors encode the predicted score or mixture weight for MDN, where darker blue corresponds to higher scores. Red circles represent the MDN's predicted variance for each Gaussian (opacity reflects mixture weight), while WTA figures depict the Voronoi tessellations for predicted hypotheses. *1st row*: 16 hypotheses, *2nd row*: 49 hypotheses.

a bias towards fitting Gaussian distributions, and they are trained by minimizing the NLL loss. We observe, as expected, excellent NLL results, especially in the case of the Single Gaussian (Figure 3 top). Note that MDN is the only method for which variable scaling factors are authorized in each hypothesis, giving it an immediate advantage. Nevertheless, Voronoi-WTA still achieves on par performance in terms of EMD and NLL on non-Gaussian datasets, as long as the scaling parameter h is well-tuned. Furthermore, as MDNs do not benefit from the optimal quantization properties of the WTA-based method, they tend to obtain suboptimal quantization errors in most cases (Figure 3 bottom).

Comparison with Kernel-WTA. We validate here the choice of Voronoi-WTA instead of the more straightforward Kernel-WTA. Figure 4 provides a comparison of both methods in the case of 16 hypotheses, in terms of NLL test performance as a function of the scaling factor h. We notice the expected behavior: in the low h-value regime, Voronoi-WTA and Kernel-WTA curves coincide, but as h increases, Voronoi-WTA's performance stabilizes, while Kernel-WTA's diverges. Results using truncated uniform kernels are also plotted in dashed lines. As expected, Voronoi-WTA's performance converges to the latter's as $h \to \infty$.

Validation of Proposition 5.2. We plot at the bottom of Figure 3 both theoretical quantization errors for Voronoi-WTA and Histogram derived in Proposition 5.2. First, we can notice that there is a good match between the theoretical errors and the empirical ones for all considered datasets. This is especially true in the asymptotic regime, where the theoretical formula becomes more accurate. This validates our assumption that our underlying score-based WTA models are close to the global minimum of their training objectives.

6.4. Evaluation on UCI Regression Datasets

In Table 1, we present additional results for the UCI datasets, adhering to the experimental protocols followed by Hernández-Lobato & Adams (2015); Lakshminarayanan et al. (2017). A more comprehensive analysis of these results is deferred to Appendix C.2. This appendix includes an extended version of Table 1 and also covers results using the RMSE metric (Table 5). Note that $\dim(\mathcal{Y}) = 1$ here.

In these datasets, the scaling factor h of WTA-based models was optimized using a golden section search (Kiefer, 1953), based on the average NLL over the validation set. Here, this optimization was costly because it was carried out very precisely. As a result, the superior sensitivity of Kernel-WTA to the choice of h, when compared to Voronoi-WTA, is not expected to be visible in these results (see the optimized NLL of Voronoi-WTA and Kernel-WTA in Figure 4), particularly as K is small ($K = 5$). Future research will explore how the optimization of h at validation time may lead to a performance disparity between Voronoi-WTA and Kernel-WTA, especially in the context of a distribution shift between validation and test samples.

These results further highlight the competitiveness of WTA-based density estimators in terms of NLL against Mixture Density Networks (MDN) and Deep ensembles (Lakshminarayanan et al., 2017), especially when the dataset size is large (*e.g.,* Protein, Year, *cf.* Table 3). This finding is particularly promising given the inherent advantages of the other baselines: indeed, NLL is not directly optimized during training in WTA-based methods. Moreover, we faced stability issues when training MDN (*e.g.,* numerical overflows in log-likelihood computation), that we did not encounter with Voronoi-WTA. We hope that these results will encourage further research into the properties of these estimators.

6.5. Experimental validation with audio data

In this section, we experimentally validate our method on a real-world application, namely on the task of Sound Event Localization (SEL) (Adavanne et al., 2018b; Grumiaux et al., 2022) which involves angular localization of sound sources from input audio signals. This task is intrinsically ambigu-

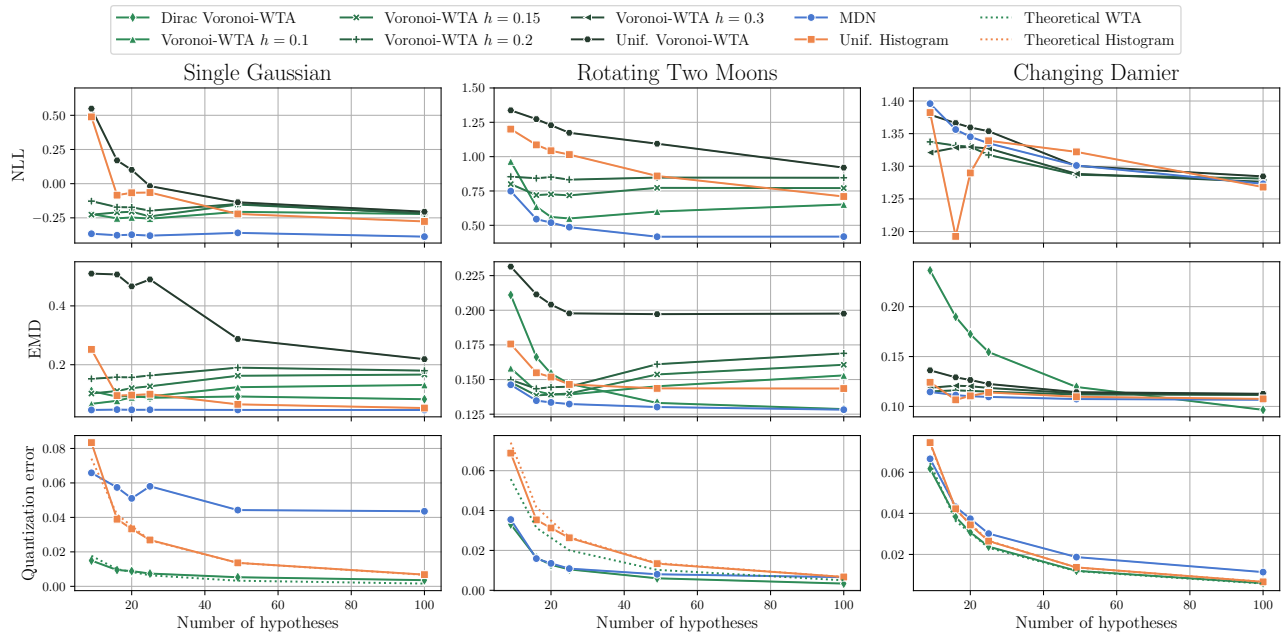

Figure 3. **Quantitative comparison.** Each column corresponds to a dataset, and each row to a different metric detailed in Section 6.1. Dotted lines correspond to theoretical quantization errors from Proposition 5.2. Dirac Voronoi-WTA corresponds to the limit when the scaling factor $h \to 0$ (9), while Unif. Voronoi-WTA is the limit when $h \to \infty$. Results are averaged over three random seeds, with standard deviations given in Appendix, Figure 7. Detailed discussion is given in Section 6.

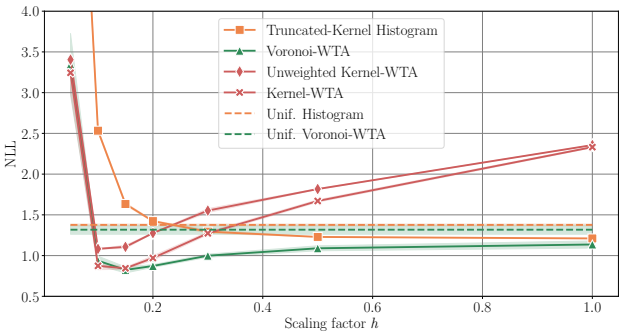

Figure 4. **Impact of the scaling factor.** Results on the dataset *Uniform to Gaussians* with 16 hypotheses, computed over three random seeds. Unweighted Kernel-WTA corresponds to (12) with fixed uniform scores $\gamma_\theta^k(x) = 1/K$. Truncated-Kernel Histogram is the standard Histogram where truncated kernels are placed on the fixed hypotheses, instead of uniform kernels (Unif. Histogram) used in Figure 3. See Appendix C.1.5 for more results.

Table 1. **UCI regression benchmark datasets comparing NLL with 5 hypotheses.** ⋆ corresponds to reported results from Lakshminarayanan et al. (2017). '–' corresponds to cases where MDN has not converged. Best results are in **bold**. ± represents the standard deviation over the official splits.

	NLL (↓)			
Datasets	Deep Ensembles⋆	MDN	K-WTA	V-WTA
Boston	**2.41 ± 0.25**	2.95 ± 0.31	**2.48 ± 0.16**	**2.48 ± 0.19**
Concrete	**3.06 ± 0.18**	3.96 ± 0.24	**3.09 ± 0.10**	**3.08 ± 0.12**
Energy	**1.38 ± 0.22**	**1.25 ± 0.25**	2.27 ± 1.22	2.22 ± 1.20
Kin8nm	**-1.20 ± 0.02**	-0.87 ± 0.05	-0.73 ± 0.03	-0.85 ± 0.05
Naval	**-5.63 ± 0.05**	**-5.47 ± 0.29**	-1.94 ± 0.00	-3.52 ± 0.38
Power	**2.79 ± 0.04**	3.02 ± 0.07	**2.81 ± 0.05**	**2.85 ± 0.06**
Protein	2.83 ± 0.02	–	**2.39 ± 0.03**	**2.42 ± 0.04**
Wine	0.94 ± 0.12	**-1.53 ± 0.76**	0.42 ± 0.18	0.37 ± 0.17
Yacht	**1.18 ± 0.21**	2.43 ± 0.72	2.23 ± 0.52	2.05 ± 0.46
Year	3.35 ± NA	–	**3.26 ± NA**	3.29 ± NA

ous, as there is spatial dispersion in the position of the sound sources to predict, either due to the sound source nature or to label noise. We considered the audio dataset ANSYN (Adavanne et al., 2018a). This dataset is generated using simulated room impulse responses so that the position of the sound sources can be considered free of noise. We injected input-dependent label noise in the source position, conditionally to the class of the input. Models are trained using this noisy data to capture this input-dependent uncertainty.

This task is challenging because it involves real-world data, label noise, and also because of the spherical geometry of the output space which calls for specific angular metrics, that are not Euclidean. Therefore, we depart from the previous theoretical and experimental setting. All models are trained using the same setup to ensure fair comparisons. We use the same evaluation framework introduced for the synthetic datasets (*cf.* Appendix C.3 for further details).

We compare in Table 2 the performance of Voronoi-WTA

Table 2. **NLL comparison of Voronoi-WTA (V-WTA) vs. Kernel-WTA (K-WTA) on audio data.** 'Hist' corresponds to the histogram with a uniform kernel as a baseline. 'Distortion' is the quantization error scaled up by 10^2. See Section 6.5 for the discussion and Appendix C.3.2 for extended analysis.

	NLL									Distortion		
h	0.3			0.5			1.0			$\emptyset$		
K	9	16	25	9	16	25	9	16	25	9	16	25
V-WTA	1.27	**1.18**	**1.15**	**1.36**	**1.24**	**1.18**	**1.57**	**1.33**	**1.22**	**0.42**	**0.26**	**0.17**
K-WTA	**1.26**	1.24	1.23	1.50	1.51	1.51	2.08	2.09	2.09	**0.42**	**0.26**	**0.17**
Hist	1.72	1.52	1.45	1.72	1.52	1.45	1.72	1.52	1.45	1.23	0.72	0.47

to Kernel-WTA using isotropic von Mises-Fisher kernels with scaling factor $h \in \{0.3, 0.5, 1.0\}$, as well as to the Histogram baseline with uniform kernels. Overall, we see that the performance of all methods tends to improve as the number of hypotheses increases, both in terms of NLL and Quantization Error. The competitive advantage of Voronoi-WTA against Kernel-WTA is confirmed, especially for large h values. We also notice that this performance gap narrows with fewer hypotheses: in a low-hypothesis regime, the Voronoi tessellation resolution is smaller, reducing the possibility of capturing local geometry through truncated kernels.

7. Related work

Conditional density estimation. Density estimation can be tackled using parametric methods (*e.g.,* Gaussian Mixture Models) or non parametric methods (*e.g.,* Kernel Density Estimation (Rosenblatt, 1956), Histograms). Mixture Density Networks Bishop (1994) is a standard deep learning extension of Gaussian Mixtures to the case of conditional densities. Their strong performances across various tasks led them to become quite popular (Zen & Senior, 2014; Li & Lee, 2019). However, MDNs notoriously suffer from numerical instabilities (Makansi et al., 2019), mode collapse (Brando Guillaumes, 2017), low contribution to the gradient of points with high predictive variance (Seitzer et al., 2022) and large biases depending on the choice of kernel (Polianskii et al., 2022). In contrast, Histogram is perhaps the simplest non-parametric alternative but is impractical in high-dimensional settings.

Multiple Choice Learning. First introduced by Guzman-Rivera et al. (2012), and adapted to deep learning by Lee et al. (2016), MCL is effective in various applications, notably in computer vision (Tian et al., 2019; Garcia et al., 2021). It suffers from two main drawbacks: hypotheses collapse and overconfidence. Solutions for the first problem include top-n update rules (Makansi et al., 2019) or allowing a small amount of gradient flow to all hypotheses (Rupprecht et al., 2017). The second problem has been solved by the introduction of scoring models (Lee et al., 2017). This approach has recently allowed for a probabilistic view of MCL (Letzelter et al., 2023). However, MCL predictions are discrete by design. One purpose of the current work is to extend MCL to density estimation, *e.g.,* for improving the evaluation of such models.

Geometry of the Voronoi tessellations. Centroidal Voronoi tessellations (Lloyd, 1982) are widely used for clustering, vector quantization (Gersho, 1979), and shape approximation (Du et al., 2003). Its popularity stems from training stability, and theoretical properties that have been extensively studied (Du et al., 1999), especially in the context of optimal quantization (Zador, 1982). This method has been used to build continuous density estimators based on uniform (Okabe et al., 2009) or Gaussian (Polianskii et al., 2022) distributions. The additional challenges raised by this continuous extension, such as volume estimation in high dimensional settings, or density discontinuity at the cell boundaries, have been discussed in the literature (Polianskii et al., 2022; Marchetti et al., 2023). However, none of these methods have yet been extended to the conditional setting, which is the topic of our work.

8. Limitations

Voronoi-WTA uses the WTA training scheme to estimate the inherent uncertainty in data. However, this approach has limitations and may achieve suboptimal performance. Recent studies have highlighted the sensitivity of WTA initialization in certain scenarios (Makansi et al., 2019; Narayanan et al., 2021). Exploring theoretically grounded solutions to address these issues, such as in Arthur (2007), could be a promising direction for future research. Additionally, there is no current evidence that the model can assess its own prediction confidence, such as in detecting out-of-distribution samples. Enhancing WTA learners with model uncertainty quantification could expand the abilities of WTA learners.

9. Conclusion

In this paper, we introduced *Voronoi-WTA*, a novel conditional density estimator. Voronoi-WTA is a probabilistic extension of traditional WTA learners, leveraging the advantageous geometric properties of the WTA training scheme. Notably, Voronoi-WTA demonstrates greater resilience to the choice of scaling factor h compared to the more straightforward Kernel-WTA. We support our claims with mathematical derivations, discussing the asymptotic performance as the number of hypotheses increases. Both theoretical analysis and experimental comparisons against several baselines highlight the strengths of our estimator. The application of our estimator to more realistic datasets opens up broad possibilities for future work.

Acknowledgements

This work was funded by the French Association for Technological Research (ANRT CIFRE contract 2022-1854) and Hi! PARIS through their PhD in AI funding programs. We are grateful to the reviewers for their insightful comments.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13 (1):34–48, 2018a.

Adavanne, S., Politis, A., and Virtanen, T. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *EUSIPCO*, pp. 1462–1466. IEEE, 2018b.

Adavanne, S., Politis, A., and Virtanen, T. Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network. *arXiv preprint arXiv:1904.12769*, 2019.

Aggarwal, A., Deshpande, A., and Kannan, R. Adaptive sampling for k-means clustering. In *APPROX*, pp. 15–28. Springer, 2009.

Arthur, D. K-means++: The advantages if careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

Bishop, C. M. Mixture density networks. 1994.

Blanchard, P., Higham, D. J., and Higham, N. J. Accurate computation of the log-sum-exp and softmax functions. *arXiv preprint arXiv:1909.03469*, 2019.

Blömer, J., Lammersen, C., Schmidt, M., and Sohler, C. Theoretical analysis of the k-means algorithm–a survey. *Algorithm Engineering: Selected Results and Surveys*, pp. 81–116, 2016.

Bourne, D. P. and Roper, S. M. Centroidal power diagrams, lloyd's algorithm, and applications to optimal location problems. *SIAM Journal on Numerical Analysis*, 53(6): 2545–2569, 2015.

Brando Guillaumes, A. Mixture density networks for distribution and uncertainty estimation. Master's thesis, Universitat Politècnica de Catalunya, 2017.

Cui, H., Radosavljevic, V., Chou, F.-C., Lin, T.-H., Nguyen, T., Huang, T.-K., Schneider, J., and Djuric, N. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *ICRA*, pp. 2090–2096. IEEE, 2019.

Devroye, L., Györfi, L., Lugosi, G., and Walk, H. On the measure of voronoi cells. *Journal of Applied Probability*, 54(2):394–408, 2017.

Du, Q., Faber, V., and Gunzburger, M. Centroidal voronoi tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, 1999.

Du, Q., Gunzburger, M. D., and Ju, L. Constrained centroidal voronoi tessellations for surfaces. *SIAM Journal on Scientific Computing*, 24(5):1488–1506, 2003.

Dua, D. and Graff, C. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

Emelianenko, M., Ju, L., and Rand, A. Nondegeneracy and weak global convergence of the lloyd algorithm in $\mathbb{R}^d$. *SIAM Journal on Numerical Analysis*, 46(3):1423–1441, 2008.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pp. 1050–1059. PMLR, 2016.

Garcia, N. C., Bargal, S. A., Ablavsky, V., Morerio, P., Murino, V., and Sclaroff, S. Distillation multiple choice learning for multimodal action recognition. In *WACV*, pp. 2755–2764, 2021.

Gersho, A. Asymptotically optimal block quantization. *IEEE Transactions on information theory*, 25(4):373–380, 1979.

Graf, S. and Luschgy, H. *Foundations of quantization for probability distributions*. Springer, 2007.

Graf, S., Luschgy, H., and Pagès, G. Distortion mismatch in the quantization of probability measures. *ESAIM: Probability and Statistics*, 12:127–153, 2008.

Gramacki, A. *Nonparametric kernel density estimation and its computational aspects*, volume 37. Springer, 2018.

Graves, A. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Grumiaux, P.-A., Kitić, S., Girin, L., and Guérin, A. A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, 152(1):107–151, 2022.

Guzman-Rivera, A., Batra, D., and Kohli, P. Multiple choice learning: Learning to produce multiple structured outputs. In *NeurIPS*, volume 25, 2012.

Han, X., Zheng, H., and Zhou, M. Card: Classification and regression diffusion models. In *NeurIPS*, volume 35, pp. 18100–18115, 2022.

Hernández-Lobato, J. M. and Adams, R. Probabilistic back-propagation for scalable learning of bayesian neural networks. In *ICML*, pp. 1861–1869. PMLR, 2015.

Hjorth, L. U. and Nabney, I. T. Regularisation of mixture density networks. In *ICANN*, volume 2, pp. 521–526. IET, 1999.

Iacobelli, M. Asymptotic quantization for probability measures on riemannian manifolds. *ESAIM: Control, Optimisation and Calculus of Variations*, 22(3):770–785, 2016.

Imani, E. and White, M. Improving regression performance with distributional losses. In *ICML*, pp. 2157–2166. PMLR, 2018.

Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 133, pp. 1381–1382, 1942.

Kiefer, J. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3): 502–506, 1953.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, volume 30, 2017.

Lee, K., Hwang, C., Park, K., and Shin, J. Confident multiple choice learning. In *ICML*, pp. 2014–2023. PMLR, 2017.

Lee, S., Purushwalkam Shiva Prakash, S., Cogswell, M., Ranjan, V., Crandall, D., and Batra, D. Stochastic multiple choice learning for training diverse deep ensembles. In *NeurIPS*, volume 29, 2016.

Letzelter, V., Fontaine, M., Chen, M., Pérez, P., Essid, S., and Richard, G. Resilient multiple choice learning: A learned scoring scheme with application to audio scene analysis. In *NeurIPS*, volume 36, 2023.

Li, C. and Lee, G. H. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *CVPR*, pp. 9887–9895, 2019.

Lloyd, S. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2018.

Makansi, O., Ilg, E., Cicek, O., and Brox, T. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *CVPR*, pp. 7144–7153, 2019.

Marchetti, G. L., Polianskii, V., Varava, A., Pokorny, F. T., and Kragic, D. An efficient and continuous voronoi density estimator. In *AISTATS*, pp. 4732–4744. PMLR, 2023.

Messaoud, S., Forsyth, D., and Schwing, A. G. Structural consistency and controllability for diverse colorization. In *ECCV*, pp. 596–612, 2018.

Narayanan, S., Moslemi, R., Pittaluga, F., Liu, B., and Chandraker, M. Divide-and-conquer for lane-aware diverse trajectory prediction. In *CVPR*, pp. 15799–15808, 2021.

Newman, D. The hexagon theorem. *IEEE Transactions on information theory*, 28(2):137–139, 1982.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. Spatial tessellations: concepts and applications of voronoi diagrams. 2009.

Pagès, G. and Printems, J. Optimal quadratic quantization for numerics: the gaussian case. *Monte Carlo Methods Appl.*, 9(2):135–165, 2003.

Parzen, E. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3): 1065–1076, 1962.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, volume 32, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Polianskii, V. and Pokorny, F. T. Voronoi boundary classification: A high-dimensional geometric approach via weighted monte carlo integration. In *ICML*, pp. 5162–5170. PMLR, 2019.

Polianskii, V., Marchetti, G. L., Kravberg, A., Varava, A., Pokorny, F. T., and Kragic, D. Voronoi density estimator for high-dimensional data: Computation, compactification and convergence. In *UAI*, pp. 1644–1653. PMLR, 2022.

Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pp. 832–837, 1956.

Rupprecht, C., Laina, I., DiPietro, R., Baust, M., Tombari, F., Navab, N., and Hager, G. D. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *ICCV*, pp. 3591–3600, 2017.

Sabin, M. and Gray, R. Global convergence and empirical consistency of the generalized lloyd algorithm. *IEEE Transactions on information theory*, 32(2):148–155, 1986.

Schymura, C., Bönninghoff, B., Ochiai, T., Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., and Kolossa, D. Pilot: Introducing transformers for probabilistic sound event localization. In *Interspeech*, pp. 2117–2121. ISCA, 2021a.

Schymura, C., Ochiai, T., Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., and Kolossa, D. Exploiting attention-based sequence-to-sequence architectures for sound event localization. In *EUSIPCO*, pp. 231–235. IEEE, 2021b.

Seitzer, M., Tavakoli, A., Antic, D., and Martius, G. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *ICLR*, 2022.

Tian, K., Xu, Y., Zhou, S., and Guan, J. Versatile multiple choice learning and its application to vision computing. In *CVPR*, pp. 6349–6357, 2019.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NeurIPS*, volume 30, 2017.

Weisstein, E. W. Sphere point picking. 2002.

Zador, P. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982.

Zen, H. and Senior, A. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *ICASSP*, pp. 3844–3848. IEEE, 2014.

Organization of the Appendix

The Appendix is organized as follows. Appendix A outlines the notations employed. Appendix B presents the theoretical results of the paper, detailing the background in Appendix B.1, demonstrating how our estimator performs distribution estimation in Appendix B.2, and discussing its geometric properties in Appendix B.3. Appendix C covers the experimental details and design choices, including the synthetic data experiments in Appendix C.1, the UCI regression benchmark in Appendix C.2 and the audio experiments in Appendix C.3.

A. Notations

Let $\mathcal{X}$ be the set of possible inputs, and by $\mathcal{Y}$ the set of possible targets. We assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite-dimensional real vector spaces, and we note d as the dimension of $\mathcal{Y}$. Given K hypotheses, let $f_\theta = (f_\theta^1, \dots, f_\theta^K) \in \mathcal{F}(\mathcal{X}, \mathcal{Y}^K)$ and $\gamma_\theta = (\gamma_\theta^1, \dots, \gamma_\theta^K) \in \mathcal{F}(\mathcal{X}, \Delta_K)$ be the predictions and scoring models, where $\Delta_K = \{p \in [0,1]^K \quad \sum_{k=1}^K p_k = 1\}$ is the simplex on $\mathbb{R}^K$. These models are described by parameters θ. When not necessary, we omit this dependency by writing $z_k = f_\theta^k$ and $\gamma_k = \gamma_\theta^k$. For a given input $x \in \mathcal{X}$ the set of predictions $\{f_\theta^k(x)\}_{k \in [\![1,K]\!]}$ induces a Voronoi tessellation of the target space $\mathcal{Y}$. We note $\mathcal{Y}_\theta^k(x)$, or alternatively $\mathcal{Y}^k(x)$, the Voronoi cell k, and by $z_k(x) = f_\theta^k(x)$ its generator. Recall that

$$\mathcal{Y}^k(x) = \{y \in \mathcal{Y} \mid \ell(y, z_k(x)) < \ell(y, z_l(x)), \forall l \neq k\} \ ,$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is the underlying loss used, for instance the L^2 loss $\ell(\hat{y}, y) = \|\hat{y} - y\|^2$, denoting by $\|\cdot\|$ the Euclidean norm.

We will drop the dependency on x when the context is clear, thus referring to $\mathcal{Y}^k$ and z_k. Conversely, when we study asymptotic properties that depend on the number of hypotheses K, we will emphasize this dependency by writing $\mathcal{Y}_K^k$ and z_K^k. Additionally, when $\mathcal{Y}$ is a d-dimensional cube, it can be partitioned into a regular grid. We note $\mathcal{G}_k$ the cubes of this grid, and by g_k the center of each cube. Note that this is a special case of Voronoi tessellation.

We note the set sum by $E + F = \{e + f, e \in E, f \in F\}$, and the border of a set E by $\partial E = \bar{E} \backslash \mathring{E}$ where $\bar{E}, \mathring{E}$ represents the closure and interior of E respectively. $B(0, r)$ is the ball of radius r with center 0 associated to $\|\cdot\|$. We note $|E|$ the cardinal of a set E, and by $\Delta(E) = \sup_{x,y \in E} \|x - y\|$ its diameter.

We note $\mathbb{P}$ a data distribution over $\mathcal{X} \times \mathcal{Y}$, λ the Lebesgue measure, δ_y the Dirac measure centered on y, $\mathbb{1}$ the indicator function, $\mathcal{U}$ the uniform distribution, and $\mathcal{N}(a, b)$ the normal distribution with mean a and variance b. We will always assume that $\mathbb{P}$ admits a probability density function $\rho(x, y)$. We denote $\mathbb{P}_x$ and ρ_x as the distribution and density, respectively, conditional on x. If p and q denote two densities over a domain $\mathcal{D}$, we define the Kullback–Leibler divergence as

$$\mathrm{KL}(p||q) = \int_{\mathcal{D}} \log\left(\frac{p(x)}{q(x)}\right) p(x) \mathrm{d}x \, .$$

When the two distributions are discrete, for instance with a support of size K, we will write

$$\mathrm{KL}_{k \in [\![1,K]\!]}(p_k \mid\mid q_k) = \sum_{k=1}^{K} \log\left(\frac{p_k}{q_k}\right) p_k \, .$$

For scalars $a, b \in (0, 1)$, we define the binary cross entropy (BCE) as

$$\mathrm{BCE}(a, b) = -a \log(b) - (1 - a) \log(1 - b) \, ,$$

adopting the convention that $0 \log 0 = 0$.

In the following, we define training objectives, which are functions of model parameters θ, using the notation $\mathcal{L} \triangleq \mathcal{L}(\theta)$. For a specific model M, the training objective is denoted by $\mathcal{L}_{\mathrm{M}}(\theta)$. The single-sample version of this objective, which we will denote as $\mathcal{L}^{\mathrm{M}}(\theta)$ for brevity, is expressed for individual data points (x, y) as $\mathcal{L}^{\mathrm{M}}(\theta, x, y)$. This single-sample loss contributes to the overall objective $\mathcal{L}_{\mathrm{M}}(\theta)$ through integration over the data distribution $\rho(x, y)$, with $\mathcal{L}_{\mathrm{M}}(\theta) = \int_{\mathcal{X} \times \mathcal{Y}} \mathcal{L}^{\mathrm{M}}(\theta, x, y) \rho(x, y) \mathrm{d}x \mathrm{d}y$.

B. Theoretical results

The estimator Voronoi-WTA (V-WTA) introduced in this paper, has two main advantages: 1) it accurately estimates the data distribution, and 2) the centroids, aligning with the optimal hypotheses according to Proposition 2.1, preserve the geometry of the data distribution. As a result, this method effectively combines the strengths of Mixture density networks and Winner-takes-all models. In this section, we will study these two claims along two main axes: convergence in distribution, and asymptotic quantization risk.

The section is organized as follows. We will first introduce the necessary definitions as well as our working hypotheses, then focus on distribution estimation, and finally study the geometrical properties of our proposed algorithm.

B.1. Theoretical setup

B.1.1. Background

We are concerned with various estimators of the conditional distribution $\mathbb{P}_x$, and study their convergence. We will make use of the Portmanteau lemma (Van der Vaart, 2000) and define weak convergence as follows.

Definition (Weak convergence)**.** *We say that a sequence of measures $(\mathbb{P}_K)_{K\in\mathbb{N}}$ converges weakly towards a measure $\mathbb{P}$, and we write $\mathbb{P}_K \underset{K\to+\infty}{\rightharpoonup} \mathbb{P}$, if $\mathbb{P}_K(E) \underset{K\to+\infty}{\longrightarrow} \mathbb{P}(E)$ for all measurable E satisfying $\mathbb{P}(\partial E) = 0$.*

We also study the convergence of sequences with finite support. In this context, we will often discuss uniform convergence, which we define below.

Definition (Uniform convergence)**.** *Let $(u_{K,k})_{(K,k)\in\mathbb{N}^2}$ denote a sequence such that $(u_{K,k})_{k\in\mathbb{N}}$ has finite support for each K. We will say that u converges uniformly towards $(v_k)_{k\in\mathbb{N}}$ if $\max_{k\in\mathbb{N}}\|u_{K,k} - v_k\| \underset{K\to+\infty}{\longrightarrow} 0$. In particular, u vanishes uniformly if $\max_{k\in\mathbb{N}}\|u_{K,k}\| \underset{K\to+\infty}{\longrightarrow} 0$.*

In what follows, we will extensively use Zador's theorem (Zador, 1982), a powerful result on the asymptotic distribution of the centroids resulting from optimal quantization, which we recall below (see Graf et al. (2008), Equation 2.3, or Iacobelli (2016), Theorem 1.3, for a more general version). This theorem will allow us to derive asymptotic properties of Winner-takes-all models.

Theorem B.1 (Zador theorem)**.** *Let $\mathbb{P} = \rho\,\mathrm{d}y$ be a Lebesgue-dominated probability measure on a compact subset $\mathcal{Y}$ of $\mathbb{R}^d$. Define the optimal quantization risk*

$$\mathcal{R}_K(\mathbb{P}) = \inf_{\mathcal{Z}\subset\mathcal{Y}:|\mathcal{Z}|\leq K} \int_{\mathcal{Y}} \min_{z\in\mathcal{Z}} \|y-z\|^2 \rho(y)\mathrm{d}y\,,$$

and the asymptotic risk for the uniform distribution $J_d = \inf_K K^{2/d}\mathcal{R}_K(\mathcal{U}([0,1]^d)$. Then

$$\lim_{K\to+\infty} K^{2/d}\mathcal{R}_K(\mathbb{P}) = J_d \left(\int_{\mathcal{Y}} \rho^{d/(d+2)}\mathrm{d}y\right)^{(d+2)/d}.$$

In addition, if $\mathcal{Z}$ minimizes the risk $\mathcal{R}_K(\mathbb{P})$, then

$$\frac{1}{K}\sum_{z\in\mathcal{Z}} \delta_z \underset{K\to\infty}{\rightharpoonup} \frac{\rho^{d/(d+2)}}{\int_{\mathcal{Y}} \rho^{d/(d+2)}(x)dx}\mathrm{d}y\,.$$

The constant J_d can be computed for simple cases ($J_1 = \frac{1}{12}$ and $J_2 = \frac{5}{18\sqrt{3}}$ (Newman, 1982)) and can be approximated for large d by $J_d \sim \dfrac{d}{2\pi e}$ (Pagès & Printems, 2003; Graf & Luschgy, 2007).

B.1.2. ESTIMATORS

Using these notations we can define several estimators of the conditional distribution $\mathbb{P}_x$, for each $x \in \mathcal{X}$:

$$\mathbb{P}_x^{\text{MDN}} : E \mapsto \sum_{k=1}^{K} \pi_k(x)\, \mathcal{N}\left(E; \mu_k(x), \Sigma_k(x)\right) \tag{17}$$

$$\mathbb{P}_x^{\text{H}} : E \mapsto \sum_{k=1}^{K} \gamma_k(x) \frac{\lambda\left(E \cap \mathcal{G}_k\right)}{\lambda\left(\mathcal{G}_k\right)} \tag{18}$$

$$\mathbb{P}_x^{\text{D-WTA}} : E \mapsto \sum_{k=1}^{K} \gamma_k(x) \delta_{z_k}(E) \tag{19}$$

$$\mathbb{P}_x^{\text{U-WTA}} : E \mapsto \sum_{k=1}^{K} \gamma_k(x) \frac{\lambda\left(E \cap \mathcal{Y}^k\right)}{\lambda\left(\mathcal{Y}^k\right)} \tag{20}$$

$$\mathbb{P}_x^{\text{K-WTA}} : E \mapsto \sum_{k=1}^{K} \gamma_k(x) K_h(z_k(x), E) \tag{21}$$

$$\mathbb{P}_x^{\text{V-WTA}} : E \mapsto \sum_{k=1}^{K} \gamma_k(x) \frac{K_h\left(z_k(x), E \cap \mathcal{Y}^k\right)}{K_h\left(z_k(x), \mathcal{Y}^k\right)}, \tag{22}$$

where $h \in \mathbb{R}_+^*$ is the scaling factor of K_h and $K_h(z_k(x), E) \triangleq \int_E K_h(z_k(x), y)\mathrm{d}y$.

Note that we obtain the variants of the original Dirac estimator $\mathbb{P}_x^{\text{D-WTA}}$ (Letzelter et al., 2023) by changing the Dirac kernel to a uniform kernel ($\mathbb{P}_x^{\text{U-WTA}}$), the kernel K_h ($\mathbb{P}_x^{\text{K-WTA}}$), or its truncated version ($\mathbb{P}_x^{\text{V-WTA}}$).

When there is no ambiguity, we will refer to these estimators by $\hat{\mathbb{P}}_x$, and their density by $\hat{\rho}_x$ (when it exists).

B.1.3. TRAINING OBJECTIVES

We recall that Winner-takes-all models are trained with two objectives: a quantization objective optimizing the position of the hypotheses $z_k(x)$ and a scoring objective enforcing that $\gamma_k(x)$ accurately estimates the probability $\mathbb{P}(\mathcal{Y}^k(x))$ of each Voronoi cell of the tessellation. induced by the hypotheses. More specifically,

$$\mathcal{L}_{\text{centroid}}(\mathcal{Z}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \min_{z \in \mathcal{Z}_x} \ell(z, y) \rho(x, y) \mathrm{d}x\mathrm{d}y\,, \tag{23}$$

$$\mathcal{L}_{\text{scoring}}(\gamma) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \sum_{k=1}^{K} \mathrm{BCE}\left[\mathbb{1}\left[y \in \mathcal{Y}^k(x)\right], \gamma_k(x)\right] \rho(x, y) \mathrm{d}x\mathrm{d}y\,, \tag{24}$$

where $\mathcal{Z} : x \mapsto \mathcal{Z}_x = \{z_k(x)\}_{k \in [\![1,K]\!]} \subset \mathcal{Y}$ and $\gamma : x \mapsto (\gamma_k(x))_{k \in [\![1,K]\!]} \in \Delta_K$.

B.1.4. ASSUMPTIONS

Throughout our analysis, we will often use the following assumptions.

Assumption B.2 (Boundedness)**.** *The set of possible outputs $\mathcal{Y}$ is compact.*

Assumption B.3 (Positivity)**.** *The data probability density function (PDF) ρ satisfies* $\inf_{(x,y) \in \mathcal{X} \times \mathcal{Y}} [\rho(x, y)] > 0$.

Assumption B.4 (Lipschitz)**.** *The conditional data PDF ρ_x is L-lipschitz for each $x \in \mathcal{X}$.*

Assumption B.5 (Optimality)**.** *The Winner-Takes-All algorithm has converged toward a global minimum of its centroid objective*

$$\min_{\mathcal{Z}} \int_{\mathcal{X}} \int_{\mathcal{Y}} \min_{z \in \mathcal{Z}_x} \ell(z, y) \rho(x, y) \mathrm{d}x\mathrm{d}y\,, \tag{25}$$

and its scoring objective (noting $x \mapsto (\mathcal{Y}^k(x))_{k\in[\![1,K]\!]}$ the resulting optimal voronoi tesselation map)

$$\min_{\gamma} \int_{\mathcal{X}} \int_{\mathcal{Y}} \sum_{k=1}^{K} \mathrm{BCE}\left[\mathbb{1}\left[y \in \mathcal{Y}^k(x)\right], \gamma_k(x)\right] \rho(x,y)\mathrm{d}x\mathrm{d}y\,. \tag{26}$$

An empirical discussion of these assumptions is given in Appendix B.4.

B.2. Distribution estimation

B.2.1. Unbiased estimators

The first interesting property of WTA is that its scoring model is an unbiased estimator of the Voronoi cell's probability mass

Proposition B.6. *Under Assumption B.5, we have*

$$\forall x \in \mathcal{X}, \quad \forall k \in [\![1,K]\!], \quad \gamma_k(x) = \mathbb{P}_x(\mathcal{Y}^k(x))\,.$$

This observation is key to establishing other interesting properties of WTA. Note that it is independent of the kernel choice, so it applies to all variants of WTA.

Proof. The scoring objective will be minimal when the integrand of Equation 26 is minimal for each $x \in \mathcal{X}$. Looking only at the integrand, we can write the following.

$$\begin{aligned}\int_{\mathcal{Y}} \sum_{k=1}^{K} \mathrm{BCE}\left[\mathbb{1}[y \in \mathcal{Y}^k(x)], \gamma_k(x)\right] \rho_x(y)\mathrm{d}y &= -\sum_{k=1}^{K} \int_{\mathcal{Y}^k(x)} \log(\gamma_k(x))\rho_x(y)\mathrm{d}y + \int_{\mathcal{Y}\setminus\mathcal{Y}^k(x)} \log(1-\gamma_k(x))\rho_x(y)\mathrm{d}y \\ &= -\sum_{k=1}^{K} \log(\gamma_k(x))\mathbb{P}(\mathcal{Y}^k(x)) + \log(1-\gamma_k(x))(1-\mathbb{P}(\mathcal{Y}^k(x)))\end{aligned}$$

For each k in the sum, we recognize a binary cross-entropy between $\gamma_k(x)$ and $\mathbb{P}_x(\mathcal{Y}^k(x))$, which is minimal when the two terms are equal. □

Therefore all truncated estimators are themselves unbiased estimators of the Voronoi cell's probability mass. We refer to this as the cell-scoring property, defined in (11).

Proposition B.7. *Under Assumption B.5, all estimators except $\mathbb{P}_x^{\mathrm{MDN}}$ and $\mathbb{P}_x^{\mathrm{K-WTA}}$ satisfy*

$$\forall x \in \mathcal{X}, \quad \forall k \in [\![1,K]\!], \quad \hat{\mathbb{P}}(\mathcal{Y}^k(x)) = \mathbb{P}(\mathcal{Y}^k(x))\,.$$

Proof. Corollary of Proposition B.6. □

B.2.2. Interpretation of the negative log-likelihood

The negative log-likelihood (NLL) is a useful quantity for measuring the accuracy of a density estimator. For instance, Mixture Density Networks minimize the NLL during training. Score-based WTA models are not trained to directly minimize NLL. The following result states that, in the case of uniform-kernel estimators, the scoring objective and the NLL are minimized when high-density zones of the target space $\mathcal{Y}$ are assigned to smaller Voronoi cells in volume.

Proposition B.8. *Under Assumption B.5, the estimator $\mathbb{P}_x^{\mathrm{U-WTA}}$ conditional negative log-likelihood satisfies for each $x \in \mathcal{X}$:*

$$\mathrm{NLL}(\mathbb{P}_x^{\mathrm{U-WTA}}, \mathbb{P}_x) = -\sum_{k=1}^{K} \log \frac{\mathbb{P}_x(\mathcal{Y}^k(x))}{\lambda\left(\mathcal{Y}^k(x)\right)}\, \mathbb{P}_x(\mathcal{Y}^k(x)) \triangleq -\mathrm{KL}_{k\in[\![1,K]\!]}\left[\mathbb{P}_x(\mathcal{Y}^k(x)) \;||\; \frac{\lambda(\mathcal{Y}^k(x))}{\mathrm{vol}(\mathcal{Y})}\right] + \log \mathrm{vol}(\mathcal{Y})\,. \tag{27}$$

From (27), we see that minimizing the NLL with constant volume $\mathrm{vol}(\mathcal{Y})$ requires strategic placement of hypotheses. Specifically, the probabilities $\mathbb{P}_x(\mathcal{Y}^k(x))$ should be high in regions where the relative volume $\frac{\lambda(\mathcal{Y}^k(x))}{\mathrm{vol}(\mathcal{Y})}$ is low.

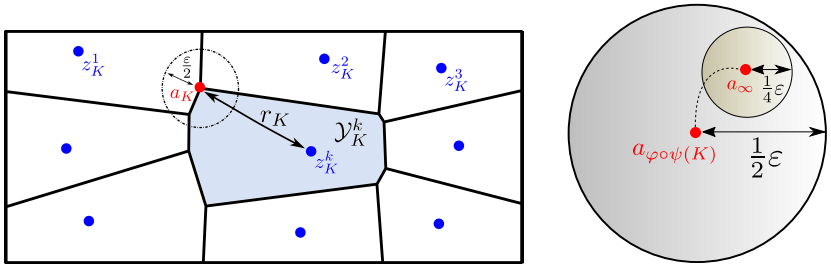

Figure 5. Illustration of the proof of Proposition B.9. On the left, we show that a_K is exactly r_K apart from its closest centroid. On the right, we illustrate the sequence $a_{\varphi\circ\psi(K)}$, from which we define B_∞.

Proof. We note $\hat{\rho}_x$ the density of the estimator $\mathbb{P}_x^{\text{U-WTA}}$. By definition (35), $\text{NLL}(\hat{\rho}_x, \rho_x) \triangleq -\int_{\mathcal{Y}} \log(\hat{\rho}_x(y))\rho_x(y)\mathrm{d}x\mathrm{d}y$. We can write

$$\begin{aligned}
\int_{\mathcal{Y}} \log(\hat{\rho}_x(y))\rho_x(y)\mathrm{d}y &\triangleq \sum_{k=1}^{K} \int_{\mathcal{Y}^k(x)} \log \frac{\gamma_k(x)}{\lambda\left(\mathcal{Y}^k(x)\right)} \rho(x,y)\mathrm{d}y \\
&= \sum_{k=1}^{K} \log \frac{\mathbb{P}_x(\mathcal{Y}^k(x))}{\lambda\left(\mathcal{Y}^k(x)\right)}\, \mathbb{P}_x(\mathcal{Y}^k(x)) \quad \text{(using proposition B.6)} \\
&= \text{KL}_{k\in[\![1,K]\!]}\left[\mathbb{P}_x(\mathcal{Y}^k(x)) \,||\, \frac{\lambda(\mathcal{Y}^k(x))}{\text{vol}(\mathcal{Y})}\right] - \log(\text{vol}(\mathcal{Y}))\left(\sum_{k=1}^{K} \mathbb{P}_x(\mathcal{Y}^k(x))\right).
\end{aligned} \tag{28}$$

□

B.2.3. Convergence of the estimators

We now turn to the main result of this section: the estimators $\mathbb{P}_x^{\text{D-WTA}}$, $\mathbb{P}_x^{\text{U-WTA}}$, and $\mathbb{P}_x^{\text{V-WTA}}$ converge in distribution towards the true data distribution $\mathbb{P}_x$.

This result is similar to Theorem 4.1 in Polianskii et al. (2022) about the convergence of the Compactified Voronoi Density Estimator (CVDE). Our proof mirrors the one they propose in this article, which is itself a slight reformulation of the Theorem 5.1 of Devroye et al. (2017). However, our setting differs from these two articles: the authors consider random *i.i.d.* generators $z_k \sim \mathbb{P}_x$ (similarly to Kernel Density Estimation (Rosenblatt, 1956; Gramacki, 2018)). However, this assumption is not satisfied in the context of WTA, which makes their result less relevant to our purpose. To correct this mismatch, we investigate the more realistic assumption that z_k minimizes the centroid objective (Assumption B.5). This makes our analysis more relevant in the context of Winner-Takes-All-based models. Additionally, we study the more general problem of conditional density estimation.

The proof of CVDE convergence relies on the intuitive observation that Voronoi cells' radius vanishes as the number of centroids K increases. Using the Zador theorem, we first show that this phenomenon still holds when the centroids are selected according to optimal quantization.

Proposition B.9. *Under Assumption B.2 (boundedness), B.3 (positive density) and B.5 (optimal centroids), the Voronoi cell diameter $\Delta(\mathcal{Y}_K^k)$ vanishes uniformly.*

Proof. Suppose that the diameter $\Delta(\mathcal{Y}_K^k)$ does not vanish. Infinitely often, there are some cells $\mathcal{Y}_K^k$ that have a diameter greater than some $\varepsilon > 0$. Inside these cells, there are points y that are more than $\frac{\varepsilon}{2}$ apart from the closest centroid. In other words, there are balls of radius $\frac{\varepsilon}{2}$ which contain no centroids (see Figure 5). This is in contrast with the second statement of Zador's theorem, which stipulates that the centroids z_K^k become dense in $\mathcal{Y}$ as K increases, hence a contradiction. We make this argument rigorous below.

We note the cell radius $r_K^k = \max_{y\in\mathcal{Y}_K^k} \|y - z_K^k\|$, and the maximal cell radius $r_K = \max_k r_K^k$. It is more convenient to work with the cell radius than with their diameter. Note that $\Delta(\mathcal{Y}_K^k) \leq 2r_K^k$ (triangular inequality), so that it is enough to show $r_K \underset{K\to\infty}{\longrightarrow} 0$.

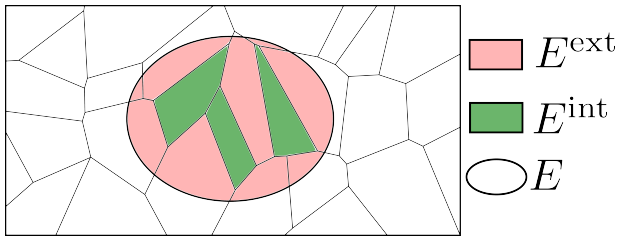

Figure 6. Illustration of the partition of $E = E^{\text{int}} \cup E^{\text{int}}$ used in the proof of Proposition B.10.

Let's assume the opposite, and let φ be a subsequence satisfying $\forall K \in \mathbb{N},\ r_{\varphi(K)} \geqslant \varepsilon$, for some $\varepsilon > 0$. We will build a ball B_∞ which contains no centroid infinitely often (see Figure 5).

Let $a_K \in \arg\max_{y \in \mathcal{Y}} \left\{ \min_{k \leq K} \left\| y - z_K^k \right\| \right\}$ be a point in the set of farthest points from their respective centroid. We can see that $B\left(a_{\varphi(K)}, \frac{\varepsilon}{2}\right)$ does not intersect any centroid at step $\varphi(K)$ (see Figure 5). Indeed, if $a_{\varphi(K)} \in \mathcal{Y}_{\varphi(K)}^k$ for some index k, then $z_{\varphi(K)}^k$ is its closest centroid and for all other centroid $z_{\varphi(K)}^l$, we have

$$\|a_{\varphi(K)} - z_{\varphi(K)}^l\| \geq \|a_{\varphi(K)} - z_{\varphi(K)}^k\| = r_{\varphi(K)} \geq \varepsilon > \frac{\varepsilon}{2}\,.$$

The sequence $(a_{\varphi(K)})_{K \in \mathbb{N}}$ is bounded (Assumption B.2), so by Bolzano–Weierstrass theorem, there is a subsequence ψ and a limiting point a_∞ such that $a_{\varphi \circ \psi(K)} \underset{K \to \infty}{\longrightarrow} a_\infty$. If K is large enough, $\|a_{\varphi \circ \psi(K)} - a_\infty\| \leq \frac{\varepsilon}{4}$ and consequently

$$B_\infty \triangleq B\left(a_\infty, \frac{\varepsilon}{4}\right) \subset B\left(a_{\varphi \circ \psi(K)}, \frac{\varepsilon}{2}\right)\,.$$

In particular, B_∞ does not intersect any centroid at each step $\varphi \circ \psi(K)$, which is exactly what we wanted to achieve.

The proportion of centroids contained in a set is given by the measure $\mathbb{P}_K = \frac{1}{K} \sum_{k=1}^K \delta_{z_K^k}$. The fact that B_∞ does not intersect any centroid can be rewritten

$$\forall K \in \varphi \circ \psi(\mathbb{N}), \quad \mathbb{P}_K\left(B_\infty\right) = \frac{1}{K} \left|\left\{k \mid z_K^k \in B_\infty\right\}\right| = 0\,.$$

Assuming optimal centroid placement, we know from Zador's theorem that

$$\mathbb{P}_K \underset{K \to \infty}{\rightharpoonup} \frac{\rho^{d/(d+2)}}{\int_{\mathcal{Y}} \rho^{d/(d+2)}(x)\mathrm{d}x}\,\mathrm{d}y \triangleq \rho_\infty\,\mathrm{d}y \triangleq \mathbb{P}_\infty\,.$$

It is clear from our hypotheses that $\inf_{y \in \mathcal{Y}} \rho_\infty(y) > 0$. We conclude with the two following contradicting observations.

$$\forall K \in \mathbb{N} \quad \mathbb{P}_{\varphi(K)}(B_\infty) = 0 \quad \Rightarrow \quad \mathbb{P}_\infty(B_\infty) = 0 \quad \text{(weak convergence)}$$

$$\mathbb{P}_\infty(B_\infty) \geq \inf_{y \in \mathcal{Y}} \rho_\infty(y) \lambda(B_\infty) > 0\,. \quad \text{(positivity)}$$

□

We can now prove the convergence of WTA-based estimators.

Proposition B.10. *Under Assumption B.2 (boundedness), B.3 (positive density) and B.5 (optimal centroids),* $\mathbb{P}_x^{\mathrm{D-WTA}}$, $\mathbb{P}_x^{\mathrm{U-WTA}}$, *and* $\mathbb{P}_x^{\mathrm{V-WTA}}$ *converge weakly towards* $\mathbb{P}$.

Proof. Let E denote any measurable such that $\lambda(\partial E) = 0$. We want to show that $\mathbb{P}_K(E) \underset{K \to +\infty}{\longrightarrow} \mathbb{P}(E)$.

If we fix K, the Voronoi tiling $\mathcal{Y}_K^k$ induces a partition of E. Accordingly, we split E as the disjoint union $E = E^{\text{int}} \cup E^{\text{ext}}$, where E^{int} denotes the Voronoi cells included in E, and E^{int} the Voronoi cells intersecting its border ∂E (see Figure 6). Using the property that the Voronoi cell diameter $\Delta(\mathcal{Y}_K^k)$ vanishes uniformly as the number of hypotheses K increases to infinity, we deduce that E^{ext} is concentrated on the border ∂E and is therefore negligible. The proof is concluded by observing that $\mathbb{P}_K$ and $\mathbb{P}$ coincide on E^{int}. We give the technical details below.

Let

$$I_K^{\text{int}} = \left\{ k \in [\![1, K]\!] \;\mid\; \mathcal{Y}_K^k \cap E \neq \varnothing,\, \mathcal{Y}_K^k \backslash E = \varnothing \right\},$$

$$I_K^{\text{ext}} = \left\{ k \in [\![1, K]\!] \;\mid\; \mathcal{Y}_K^k \cap E \neq \varnothing,\, \mathcal{Y}_K^k \backslash E \neq \varnothing \right\},$$

$E_K^{\text{int}} = \cup_{k \in I^{\text{int}}} (\mathcal{Y}_K^k \cap E)$ and $E_K^{\text{ext}} = \cup_{k \in I^{\text{ext}}} (\mathcal{Y}_K^k \cap E)$. Clearly $E = E_K^{\text{int}} \cup E_K^{\text{ext}}$ and $E_K^{\text{int}} \cap E_K^{\text{ext}} = \varnothing$.

Recall that the considered estimators have an interesting property: the estimated density of a Voronoi cell is equal to its true probability mass (Proposition B.7). Therefore, $\mathbb{P}_K$ and $\mathbb{P}$ coincide on E_K^{int}. More precisely,

$$\mathbb{P}_K(E_K^{\text{int}}) = \sum_{k \in I^{\text{int}}} \mathbb{P}_K(\mathcal{Y}_K^k) = \sum_{k \in I^{\text{int}}} \mathbb{P}\left(\mathcal{Y}_K^k\right) = \mathbb{P}\left(E_K^{\text{int}}\right). \tag{29}$$

We then refer to the maximum cell diameter by $\varepsilon_K = \max_k \Delta(\mathcal{Y}_K^k)$, and $\varepsilon_K^+ = \sup_{k \geq K} \varepsilon_k$. We know from Proposition B.9 that $\varepsilon_K \underset{K \to +\infty}{\longrightarrow} 0$.

We now show that $E_K^{\text{ext}} \subset \partial E + B(0, \varepsilon_K^+)$.

Let $k \in I_K^{\text{ext}}$. We want to show that $\mathcal{Y}_K^k$ intersects ∂E (see Figure 6). By definition, $\mathcal{Y}_K^k$ is partially inside and outside E. Therefore, we can choose $x \in \mathcal{Y}_K^k \cap E$ and $y \in \mathcal{Y}_K^k \backslash E$. By convexity of the Voronoi cells, we can see that its border ∂E will intersect the segment $[x, y]$ on a single point $y^* \in \mathcal{Y}_K^k$. Formally, let

$$t^* = \sup\{t \in [0,1] \;\mid\; (1-t)x + ty \in E\}, \quad \text{and} \quad y^* = (1 - t^*)\, x + t^* y\,.$$

We have $y^* \in \bar{E}$ because there is a sequence converging toward y^* from inside E by definition of $\sup$. Moreover, $y^* \notin E^{\text{o}}$, because there would be $t > t^*$ satisfying the constraint, by definition of open sets. Finally $y^* \in \mathcal{Y}_K^k$ by convexity of $\mathcal{Y}_K^k$. Therefore $y^* \in \partial E \cap \mathcal{Y}_K^k$.

By definition of the maximum diameter ε_K^+,

$$y^* \in \mathcal{Y}_K^k \quad \Rightarrow \quad \mathcal{Y}_K^k \subset B(y^*, \varepsilon_K^+) = y^* + B(0, \varepsilon_K^+) \subset \partial E + B(0, \varepsilon_K^+)\,.$$

Therefore, $E_K^{\text{ext}} \subset \cup_{k \in I^{\text{ext}}} \mathcal{Y}_K^k \subset \partial E + B(0, \varepsilon_K^+)$. We conclude by observing that $\mathcal{Y}_K^k$ are disjoint, and that $\mathbb{P}_K(\mathcal{Y}_K^k \cap E) \leq \mathbb{P}_K(\mathcal{Y}_K^k) = \mathbb{P}(\mathcal{Y}_K^k)$ for the considered estimators (Proposition B.7).

$$\mathbb{P}_K(E_K^{\text{ext}}) = \mathbb{P}_K(\cup_{k \in I^{\text{ext}}} (\mathcal{Y}_K^k \cap E)) \leq \mathbb{P}(\cup_{k \in I^{\text{ext}}} (\mathcal{Y}_K^k)) \leq \mathbb{P}(\partial E + B(0, \varepsilon_K^+)) \underset{K \to +\infty}{\longrightarrow} \mathbb{P}(\partial E) = 0\,. \tag{30}$$

Likewise,

$$\mathbb{P}(E_K^{\text{ext}}) \leq \mathbb{P}(\partial E + B(0, \varepsilon_K^+)) \underset{K \to +\infty}{\longrightarrow} \mathbb{P}(\partial E) = 0\,. \tag{31}$$

In conclusion,

$$|\mathbb{P}_K(E) - \mathbb{P}(E)| \leq \underbrace{|\mathbb{P}_K(E_K^{\text{int}}) - \mathbb{P}(E_K^{\text{int}})|}_{= 0 \text{ by Eq. (29)}} + \underbrace{|\mathbb{P}_K(E_K^{\text{ext}}) - \mathbb{P}(E_K^{\text{ext}})|}_{\to 0 \text{ by (30) and Eq. (31)}} \underset{K \to +\infty}{\to} 0\,.$$

□

Note that Eq. (29) does not necessarily hold for $\mathbb{P}_x^{\text{K-WTA}}$. Therefore, this proof cannot apply to this estimator. However, it applies to a large family of estimators $\hat{\mathbb{P}}$. Indeed, it is independent of the choice of kernel, as long as $\hat{\mathbb{P}}$ is an unbiased estimator of the Voronoi cell probability mass.

B.3. Geometrical properties

We will study the geometrical properties of WTA-based models through the lens of quantization risk. Our analysis will focus on the estimators $\mathbb{P}_x^{\mathrm{D-WTA}}$, $\mathbb{P}_x^{\mathrm{V-WTA}}$, and $\mathbb{P}_x^{\mathrm{K-WTA}}$, as these correspond to the cases presented in the main paper. Specifically, we will concentrate on the positions of the hypotheses, an aspect for which the same analysis applies to all three estimators. Indeed, both $\mathbb{P}_x^{\mathrm{V-WTA}}$ and $\mathbb{P}_x^{\mathrm{K-WTA}}$ retain the hypotheses positions following WTA training. For the rest of this section, we drop the dependency on x without loss of generality, to lighten the notational burden.

B.3.1. Quantization risk

We are interested in measuring the advantage of an adaptative grid of WTA. To do so, we look at the quantization risk of WTA and Histogram.

Proposition B.11. *Under Assumption B.5, we have for each $K \in \mathbb{N}$*

$$\sum_{k=1}^{K} \int_{\mathcal{Y}^k} \|z_k - y\|_2^2 \rho(y) \mathrm{d}y \leq \sum_{k=1}^{K} \int_{\mathcal{G}^k} \|g_k - y\|_2^2 \rho(y) \mathrm{d}y \,.$$

Proof. The histogram is a particular case of Voronoi tesselation. □

B.3.2. Asymptotic quantization risk

We know the asymptotic quantization risk of WTA from the Zador theorem.

Proposition B.12. *Under assumptions B.2 (boundedness) and B.5 (optimality), the asymptotic quantization risk of the estimator $\mathbb{P}_x^{\mathrm{D-WTA}}$ has the following asymptotic evolution as $K \to \infty$*

$$\mathcal{R}_K^{\mathrm{D-WTA}} = J_d \left(\int_{\mathcal{Y}} \rho^{d/(d+2)} \mathrm{d}y \right)^{(d+2)/d} \frac{1}{K^{2/d}} + o\left(\frac{1}{K^{2/d}} \right) .$$

Proof. It is a corollary of the Zador theorem, whose conditions are met given our hypotheses. □

It is also possible to compute the risk for the grid estimator $\mathbb{P}_x^{\mathrm{H}}$.

Proposition B.13. *Under assumptions B.2 (boundedness), B.3 (positivity), B.4 (lipschitz), B.5 (optimality), and assuming moreover that $\mathcal{Y} = [0, c]^d$, the quantization risk of the estimator $\mathbb{P}_x^{\mathrm{H}}$ has the following asymptotic evolution*

$$\mathcal{R}_K^{\mathrm{H}} = \frac{d}{12} \frac{c^2}{K^{2/d}} + \mathcal{O}\left(\frac{1}{K^{3/d}} \right) .$$

Proof. Consider a d-dimensional grid of $K = M^d$ points. Admitting for now that the risk over a d-dimensional cube of size $\frac{c}{M}$ is equal to $\frac{d}{12} \left(\frac{c}{M}\right)^{d+2}$, we can rewrite the risk as follows.

$$\mathcal{R}_{M^d}^{\mathrm{H}} = \sum_{k=1}^{M^d} \int_{\mathcal{Y}^k} \rho(y) \|y - z_K^k\|^2 \mathrm{d}y \approx \sum_{k=1}^{M^d} \rho(z_K^k) \int_{\mathcal{Y}^k} \|y - z_K^k\|^2 \mathrm{d}y = \frac{d}{12} \frac{c^2}{M^2} \left(\frac{c^d}{M^d} \sum_{k=1}^{M^d} \rho(z_k) \right) \approx \frac{d}{12} \frac{c^2}{M^2} \,.$$

The first approximation says that ρ is essentially constant over a cube cell V_k if K is large enough. The second approximation is a Monte Carlo integration. If we substitute M by K, we obtain the announced result.

We now justify these two approximations formally. We first compute the risk over a d-dimensional cube of side $a > 0$ centered in 0 (considering a uniform distribution for ρ).

$$\int_{[-\frac{a}{2}, \frac{a}{2}]^d} \left(x_1^2 + \cdots + x_d^2\right) \mathrm{d}x_1 \ldots \mathrm{d}x_d = d \int x_1^2 \mathrm{d}x_1 \ldots \mathrm{d}x_d = d a^{d-1} \int_{-a/2}^{a/2} x_1^2 \mathrm{d}x_1 = \frac{d}{12} a^{d+2} \,.$$

Now we compute an upper bound of the α-distortion (defined below) over the same cube. This upper bound relies on enclosing the cube in the smallest ball containing it (which has for diameter the largest diagonal of the cube $\|(a,\dots,a)-(0,\dots,0)\| = \sqrt{d}a$).

$$\begin{aligned}\int_{[-\frac{a}{2},\frac{a}{2}]^d} \left(\sqrt{x_1^2+\cdots+x_d^2}\right)^\alpha \mathrm{d}x_1\cdots\mathrm{d}x_d &\leqslant \int_{B\left(0,\sqrt{d}a/2\right)} \left(\sqrt{x_1^2+\cdots+x_d^2}\right)^\alpha \mathrm{d}x_1\cdots\mathrm{d}x_d \\ &= \int_0^{\sqrt{d}\frac{a}{2}} r^\alpha \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} r^{d-1}\mathrm{d}r \\ &= \frac{2\pi^{d/2}}{\Gamma\left(\frac{d}{2}\right)} \int_0^{\sqrt{d}\frac{a}{2}} r^{d-1+\alpha}\mathrm{d}r \\ &= \mathcal{O}\left(a^{d+\alpha}\right),\end{aligned}$$

Where $\Gamma : x \in \mathbb{R}_+^* \mapsto \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$ is the gamma function. Equipped with this result, we can prove the first approximation using the assumption that ρ is L-lipschitz.

$$\begin{aligned}\left|\mathcal{R}_K^{\mathrm{H}} - \sum_{k=1}^K \rho(z_K^k)\int_{\mathcal{Y}_K^k} \|y-z_K^k\|^2\mathrm{d}y\right| &\leqslant \sum_{k=1}^K \int_{\mathcal{Y}_K^k} |\rho(y)-\rho(z_K^k)|\,\|y-z_K^k\|^2\mathrm{d}y \\ &\leqslant L\sum_{k=1}^K \int_{\mathcal{Y}_K^k} \|y-z_K^k\|^3\mathrm{d}y \\ &= L\sum_{k=1}^K \mathcal{O}\left(\frac{1}{M^{d+3}}\right) \\ &= \mathcal{O}\left(\frac{1}{M^3}\right).\end{aligned}$$

The second approximation is similar.

$$\begin{aligned}\left|\frac{1}{M^d}\sum_{k=1}^K \rho(z_K^k) - \int_{\mathcal{Y}} \rho(y)\mathrm{d}y\right| &= \left|\sum_{k=1}^K \rho(z_K^k)\int_{\mathcal{Y}_K^k} 1\mathrm{d}y - \sum_{k=1}^N \int_{\mathcal{Y}_K^k} \rho(y)\mathrm{d}y\right| \\ &\leqslant \sum_{k=1}^K \int_{\mathcal{Y}_K^k} |\rho(y)-\rho(z_K^k)| \\ &\leqslant L\sum_{k=1}^K \int_{\mathcal{Y}_K^k} \|y-z_K^k\|\mathrm{d}y \\ &= \sum_{k=1}^K \mathcal{O}\left(\frac{1}{M^{d+1}}\right) \\ &= \mathcal{O}\left(\frac{1}{M}\right).\end{aligned}$$

Combining both, we obtain the desired result (note that $\int_{\mathcal{Y}} \rho(y)\mathrm{d}y = 1$).

$$\mathcal{R}_K^{\mathrm{H}} = \frac{d}{12}\frac{c^2}{M^2} + \mathcal{O}\left(\frac{1}{M^3}\right).$$

We can replace K in this formula.

$$\mathcal{R}_K^{\mathrm{H}} = \frac{d}{12}\frac{c^2}{K^{2/d}} + \mathcal{O}\left(\frac{1}{K^{3/d}}\right).$$

□

While the above proof was carried out with $\mathcal{Y} = [0, c]^d$ with a volume of $\lambda(\mathcal{Y}) = c^d$, one can show the more general expression:

$$\mathcal{R}_K^{\mathrm{H}} = \frac{d}{12} \frac{\lambda(\mathcal{Y})^{2/d}}{K^{2/d}} + \mathcal{O}\left(\frac{1}{K^{3/d}}\right). \tag{32}$$

Note that in the first order, the asymptotic quantization risk of the histogram does not depend on the distribution of the probability mass ρ but only on the width of its support.

Both WTA and Histogram have the same asymptotic risk $\mathcal{O}\left(K^{-\frac{2}{d}}\right)$. However, the leading constant is always larger for Histogram (for all densities ρ) for large K. This means that WTA is strictly better than Histogram even in the asymptotic regime (Proposition B.14).

Proposition B.14. *Under the assumptions of Proposition B.13, the leading constant of $\mathcal{R}_K^{\mathrm{H}}$ is greater than that of $\mathcal{R}_K^{\mathrm{D-WTA}}$ for dimension $d \in \{1, 2\}$ and for large d.*

Proof. We show that the leading constant of $\mathcal{R}_K^{\mathrm{H}}$ is greater than that of $\mathcal{R}_K^{\mathrm{WTA}}$ for large d. Using Eq. (32), we can write the following.

$$\begin{aligned}
\frac{d}{12}\lambda(\mathcal{Y})^{2/d} &> \frac{d}{2\pi e}\left(\int_{\mathcal{Y}} \mathrm{d}y\right)^{2/d} \\
&= \frac{d}{2\pi e}\left(\int_{\mathcal{Y}} \mathrm{d}y\right)^{(2+d)/d} \frac{\int_{\mathcal{Y}} \rho(y)\mathrm{d}y}{\int_{\mathcal{Y}} \mathrm{d}y} && \left(\text{since } \int_{\mathcal{Y}} \rho(y)\mathrm{d}y = 1\right) \\
&= \frac{d}{2\pi e}\left(\int_{\mathcal{Y}} \mathrm{d}y\right)^{(2+d)/d} \frac{\int_{\mathcal{Y}} \left(\rho(y)^{d/(d+2)}\right)^{(d+2)/d} \mathrm{d}y}{\int_{\mathcal{Y}} \mathrm{d}y} && (\text{since } (x^r)^{1/r} = x) \\
&\geq \frac{d}{2\pi e}\left(\int_{\mathcal{Y}} \mathrm{d}y\right)^{(2+d)/d} \left(\frac{\int_{\mathcal{Y}} \rho(y)^{d/(d+2)}\mathrm{d}y}{\int_{\mathcal{Y}} \mathrm{d}y}\right)^{(d+2)/d} && (\text{Jensen inequality applied to } x \mapsto x^{(d+2)/2}) \\
&= \frac{d}{2\pi e}\left(\int_{\mathcal{Y}} \rho(y)^{d/(d+2)}\mathrm{d}y\right)^{(d+2)/d}.
\end{aligned}$$

We recognize the leading constant of the Zador theorem on the last term (using the asymptotic equivalent $J_d \sim \frac{d}{2\pi e}$). The argument is the same for $d = 1$ (in which case $J_d = \frac{1}{12}$), and $d = 2$ (in which case $J_d = \frac{5}{18\sqrt{3}} < \frac{2}{12}$). □

B.4. Discussion of the assumptions

We discuss in this section the validity of our assumptions.

Boundedness assumption B.2. The boundedness assumption is a necessary assumption to ensure that the volume of each cell is well-defined, such as in the case of a piece-wise uniform distribution. This is a customary assumption in theoretical analysis of K-Means algorithm (Emelianenko et al., 2008), as well as in centroidal Voronoi tessellations (Du et al., 1999). Moreover, this assumption is valid in all realistic settings of pragmatic interest.

Positivity assumption B.3. The positivity assumption asserts that the data PDF is nonzero everywhere. This technical assumption is also reasonable for practical applications. For instance, if we assume that $\mathcal{X} \times \mathcal{Y}$ is bounded, modifying the distribution to $\tilde{\rho}(x, y) = (1 + \varepsilon)^{-1}(\rho(x, y) + \varepsilon)$ with a sufficiently small $\varepsilon > 0$ does not alter the experimental results, and ensures that $\tilde{\rho} > 0$ everywhere.

Lipschitz assumption B.4. The Lipschitz assumption of the data distribution was applied in Proposition B.13 and B.14. Note that no further assumption was made on the value of $L > 0$.

Optimality assumption B.5. The assumption of quantization optimality is strong. There are two main arguments for its validity:

- The convergence of WTA has not been directly studied in the literature. However, we can see WTA as a *conditional* gradient descent version of K-means (Pagès & Printems, 2003). The convergence of K-means has been extensively

studied (Sabin & Gray, 1986; Emelianenko et al., 2008; Bourne & Roper, 2015), and some results could, as further work, be ported to WTA (Pagès & Printems, 2003). Even though few results concern the convergence towards a global minimum of the quantization objective, theoretical evidence points toward the surprising effectiveness of this approach in most cases (Blömer et al., 2016).

- Our justification for Assumption B.5 is also empirical. We confirmed experimentally that the empirical quantization risk closely follows the theoretical optimal quantization risk given by Proposition 5.2. Indeed, we can see in Figure 3 that the solid and dotted green curves of the quantization risk converge in the asymptotic regime, which is when the theoretical formula is valid.

This gives us confidence that Assumption B.5 holds in practice, possibly in an approximate manner. Further research could explore in greater detail how the initialization of the Winner-Takes-All (WTA) training scheme influences the quality of the optimal solution. This investigation could build upon the findings of studies such as those by Arthur (2007) and Aggarwal et al. (2009).

C. Experimental details

C.1. Synthetic data experiments

C.1.1. Baselines.

Two standard conditional density estimation baselines were considered in our experiments: Mixture Density Networks (MDN) (Bishop, 1994) and the Histogram (Imani & White, 2018) mentioned in Section 5.2. We took care to assess all methods fairly, by using the same backbone architecture and number of hypotheses for all of them. These baselines differ however from WTA-based methods on two main axes, output format and training loss, as explained hereafter:

- *MDN* uses a multi-head neural network to directly predict the parameters of a mixture of Gaussians. It is trained using the negative log-likelihood loss:
$$\mathcal{L}^{\mathrm{MDN}}(\theta) = -\log \hat{\rho}_\theta(y \mid x)\,, \tag{33}$$
where $\hat{\rho}_\theta(y \mid x)$ is a mixture of Gaussians with parameters $\{\pi_k(x), \mu_k(x), \sigma_k(x)\}$, respectively denoting mixture weights, means and standard deviations. We considered isotropic Gaussians here. We enhanced the original MDN training scheme by incorporating the findings from Brando Guillaumes (2017) to improve the numerical stability of the training. In particular, we employed the *Log-sum-exp* trick (Blanchard et al., 2019), and modified the neural network's output to predict $(\mu, \log \sigma^2)$ rather than (μ, σ).
- *Histogram* is purely non-parametric, unlike MDN. It uses a multi-head neural network to predict scores $\gamma_\theta^k(x) \in [0, 1]$ for each point in a fixed regular grid, defining the histogram bins. It is trained through backpropagation using the following loss:
$$\mathcal{L}^{\mathrm{H}}(\theta) = -\log \gamma_\theta^{k^\star}(x) - \sum_{k \neq k^\star} \log(1 - \gamma_\theta^k(x))\,, \tag{34}$$
where $k^\star = \mathrm{argmin}_k \ell(g_k, y)$ is the bin index in which the target falls. Note that the Histogram baseline can be seen as a specific instance of hypothesis-scores architecture used in the WTA setup, where each hypothesis is static and represents the central point of a histogram bin. In the context of the synthetic data experiments of Section 6, the output space is $\mathcal{Y} = [-1, 1]^2$ and we employed a regular grid defined by row $i \in [\![1, N_{\mathrm{rows}}]\!]$ and by column $j \in [\![1, N_{\mathrm{cols}}]\!]$. For each $x \in \mathcal{X}$, if k is the hypothesis index associated to bin (i, j), we therefore have:
$$f_\theta^k(x) = \left(-1 + \left(i - \frac{1}{2}\right)\frac{2}{N_{\mathrm{rows}}}, -1 + \left(j - \frac{1}{2}\right)\frac{2}{N_{\mathrm{cols}}}\right)\,,$$
with $K = N_{\mathrm{rows}} N_{\mathrm{cols}}$. In our comparisons described in Section 6, we used $N_{\mathrm{rows}} = N_{\mathrm{cols}}$ in our experiments, except when $K = 20$ where we set $N_{\mathrm{rows}} = 5$ and $N_{\mathrm{cols}} = 4$.

C.1.2. Architectures and training details

Architectures. In each training setup with synthetic data, we employed a two-hidden-layer multilayer perceptron. Each layer contained 256 hidden units and used ReLU activation functions. In the final layer, we utilized tanh activations for the hypotheses and sigmoid activations for the scores, where applicable. The last layer is duplicated depending on the number of outputs to produce: three times the number of modes in the case of MDN (mixture coefficients, means, and variances), twice

the number of hypotheses in the WTA setup (scores and hypotheses are predicted), and the product of the number of rows and columns for the 2-dimensional histogram. Note that if the normalization of the scores is not inherently implemented in the architecture (*e.g.,* with a softmax activation), they must be normalized when computing metrics by considering $\frac{\gamma_k(x)}{\sum_k \gamma_k(x)}$.

Training details. The Adam optimizer (Kingma & Ba, 2014) was used with a constant learning rate of 0.001 in each setup. The models were trained until convergence of the training loss, using early stopping to select the checkpoint for which the validation loss was the lowest. Each of the synthetic datasets consists of $100,000$ training points, and $25,000$ validation points. Each of the models was trained for 100 epochs, with a batch size of 1024.

In each setup that involves WTA training, we used the compound loss $\mathcal{L}^{\text{WTA}} + \beta \mathcal{L}^{\text{scoring}}$ with $\beta = 1$. Note that we observed that the WTA training scheme leads to a fast convergence of the predictions $f_\theta(x)$, while the scoring heads $\gamma_\theta(x)$ are slightly slower to train. Indeed, each $\gamma_\theta^k(x)$ solves a binary classification task that evolves as the position of $f_\theta(x)$ is updated during training. Therefore, this objective is untractable at the beginning of the training, because the prediction $f_\theta(x)$ moves too quickly, and it only becomes feasible near the end of training, once the prediction has stabilized. This warrants further research on the scheduling of β during training.

C.1.3. Metrics

For assessing the quality of the predictions, we used the following metrics for each input $x \in \mathcal{X}$.

- The Negative Log-Likelihood (NLL), which assesses the probabilistic quality of the predictions

$$\text{NLL}(\hat{\rho}_x, \rho_x) = -\int_{\mathcal{Y}} \log \hat{\rho}_x(y) \rho_x(y) \mathrm{d}y \,, \tag{35}$$

  where $\hat{\rho}_x(y)$ is the estimated density, which is assumed to integrate to 1.
- The Earth Mover's Distance (EMD):

$$\text{EMD}(\hat{\rho}_x, \rho_x) = \min_{\psi \in \Psi} \sum_{y_s \sim \rho_x} \sum_{\hat{y}_k \sim \hat{\rho}_x} \psi_{s,k} \|y_s - \hat{y}_k\| \,, \tag{36}$$

  where $\psi \in \Psi$ is a transport plan belonging to the set of valid transport plans (Kantorovich, 1942).
- The Quantization Error, as defined in Pagès & Printems (2003):

$$\mathcal{R}(\mathcal{Z}) = \int_{\mathcal{Y}} \min_{z \in \mathcal{Z}} \|y - z\|^2 \, \rho_x(y) \mathrm{d}y \,. \tag{37}$$

Note that equations (35), (36) and (37) assume that the target distribution ρ_x is available. However, this is not usually the case for real-world tasks, for which typically one sample $y \sim \rho_x$ is available for each input x. In this case, the EMD loses its interpretation as a distance between distributions. Nevertheless, the NLL and Distortion errors can still be computed on an average basis over the test set, with the NLL and Quantization Errors defined as

$$\text{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log \hat{\rho}_{x_i}(y_i) \,, \tag{38}$$

$$\mathcal{R} = \frac{1}{N} \sum_{i=1}^{N} \min_{z \in \mathcal{Z}_i} \|y_i - z\|^2 \,, \tag{39}$$

where N is the number of pairs (x_i, y_i) in the test set, and $\mathcal{Z}_i = \left\{ f_\theta^l(x_i) \right\}_{l \in [\![1,K]\!]}$.

C.1.4. Evaluation details

The results of Figure 3 were computed according to the following details. Note that each evaluation was performed with $N = 2,000$ test points. The results are averaged over three random seeds (see Figure 7).

NLL computation. The NLL was computed following (38), with N the number of points on each test set, and $\hat{\rho}_x$ is for instance given in (13) in Voronoi-WTA. The Volume

$$V(f_\theta^k(x), K_h) = \int_{\mathcal{Y}_k(x)} K_h(f_\theta^k(x), \tilde{y}) \mathrm{d}\tilde{y} \,, \tag{40}$$

was computed with the normalized Gaussian kernel:

$$K_h(f_\theta^k(x), y) = \frac{1}{(2\pi)^{\frac{d}{2}} h^d} \exp\left(\frac{\|y - f_\theta^k(x)\|^2}{2h^2}\right), \tag{41}$$

where $d = \dim(\mathcal{Y})$. In particular, $d = 2$ for the synthetic data experiments and $d = 1$ for the UCI datasets experiments.

In practice, the Volume (40) can be computed efficiently, rewriting each $\tilde{y}$ in the integral as

$$y = f_\theta^k(x) + ts\,, \tag{42}$$

where $s \in \mathbb{S}^{d-1}$ is a direction on the unit sphere and $t \in [0, l_{f_\theta^k(x)}(s)]$ a scalar step. Here $l_{f_\theta^k(x)}(s)$ is the so-called *directional radius*, defined as the maximum $u \in \mathbb{R}_+$ such that $f_\theta^k(x) + us \in \mathcal{Y}_\theta^k(x)$ if it exists, and $l_{f_\theta^k(x)}(s) = \infty$ otherwise. In practice, we computed each directional radius in $\mathcal{O}(Kd)$ operations by following Equations 7 and 8 from Polianskii et al. (2022), leveraging the structure of Voronoi tesselation as detailed in Polianskii & Pokorny (2019). The case of unbounded Voronoi cells did not appear here since the output is restricted to the square $[-1, 1]^2$. As explained by Polianskii et al. (2022, Sec.3.1), (42) allows writing $V(f_\theta^k(x), K_h)$ as a double-integral in spherical coordinates:

$$V(f_\theta^k(x), y) = \int_{s \in \mathbb{S}^{d-1}} \int_{t \in \left[0, l_{f_\theta^k(x)}(s)\right]} K(\frac{t}{h}s) t^{d-1} \,\mathrm{d}t\,\mathrm{d}s\,,$$

where: $K(x) \triangleq \exp(-\frac{\|x\|^2}{2})$.

As the inner integral has a closed-form solution for the Gaussian kernel (Polianskii et al., 2022), a Monte Carlo approximation of the outer integral allows us to estimate $V(f_\theta^k(x), K_h)$:

$$V(f_\theta^k(x), y) \simeq \frac{2\pi^{\frac{d}{2}}}{N'\Gamma\left(\frac{n}{2}\right)} \sum_{j=1}^{N'} \int_{\left[0, l_{f_\theta^k(x)}(s_j)\right]} K(\frac{t}{h}s_j) t^{d-1} \,\mathrm{d}t \simeq \frac{1}{N'} \sum_{j=1}^{N'} (2\pi h^2)^{\frac{d}{2}} \bar{\gamma}\left(\frac{d}{2}, \frac{l_{f_\theta^k(x)}(s_j)^2}{2h^2}\right), \tag{43}$$

where Γ is the gamma function, N' is the number of points $\{s_j\}$ sampled on the unit sphere $\mathbb{S}^{d-1}$ (or *versors*) and $\bar{\gamma}(a, z) \triangleq \frac{1}{\Gamma(a)} \int_0^z t^{a-1} e^{-t} \,\mathrm{d}t$ is the incomplete gamma function. When $d = 2$, (43) simplifies to:

$$V(f_\theta^k(x), y) \simeq \frac{1}{N'} \sum_{j=1}^{N'} 2\pi h^2 \left(1 - \exp\left(-\frac{l_{f_\theta^k(x)}(s_j)^2}{2h^2}\right)\right).$$

In practice, we used $N' = 40$ for our experiments.

EMD computation. The EMD was computed as

$$\mathrm{EMD} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{EMD}(\hat{\rho}_{x_i}, \rho_{x_i})\,,$$

where $N = 2,000$ and $\mathrm{EMD}(\hat{\rho}_x, \rho_x)$ is defined in (36). Computing the EMD for each input requires sampling from both the predicted and target distributions, with the assumption that sampling from the target distribution is feasible. For the predicted distribution, a rejection sampling procedure was implemented. This involved initially selecting a cell k based on the distribution of scores $\left\{\gamma_\theta^l(x)\right\}$. Samples were then repeatedly drawn from the distribution $K_h(f_\theta^k(x), \cdot)$ until a sample falling within the cell $\mathcal{Y}_\theta^k(x)$ was obtained. In our experiments, we matched empirical measures by taking 500 samples each from both the predicted and target distributions. Additionally, to reduce computational complexity, especially when dealing with a large number of hypotheses, one can employ the hit-and-run sampling technique as outlined in Alg. 2 by Polianskii et al. (2022).

Theoretical curves. The theoretical curves of Figure 3 were computed according to (15) and (16). The calculation of (15) was performed through Monte-Carlo integration using $10,000$ samples across the output space for each input x (averaging over 10 inputs here), for $K \in \{9, 16, 25, 49, 100\}$. The target density ρ is explicitly defined in the *Changing Damier* and *Single Gaussian* datasets, while in the *Rotating Moons* dataset, it was approximated through Kernel Density Estimation (with Gaussian kernel and bandwidth of 0.2).

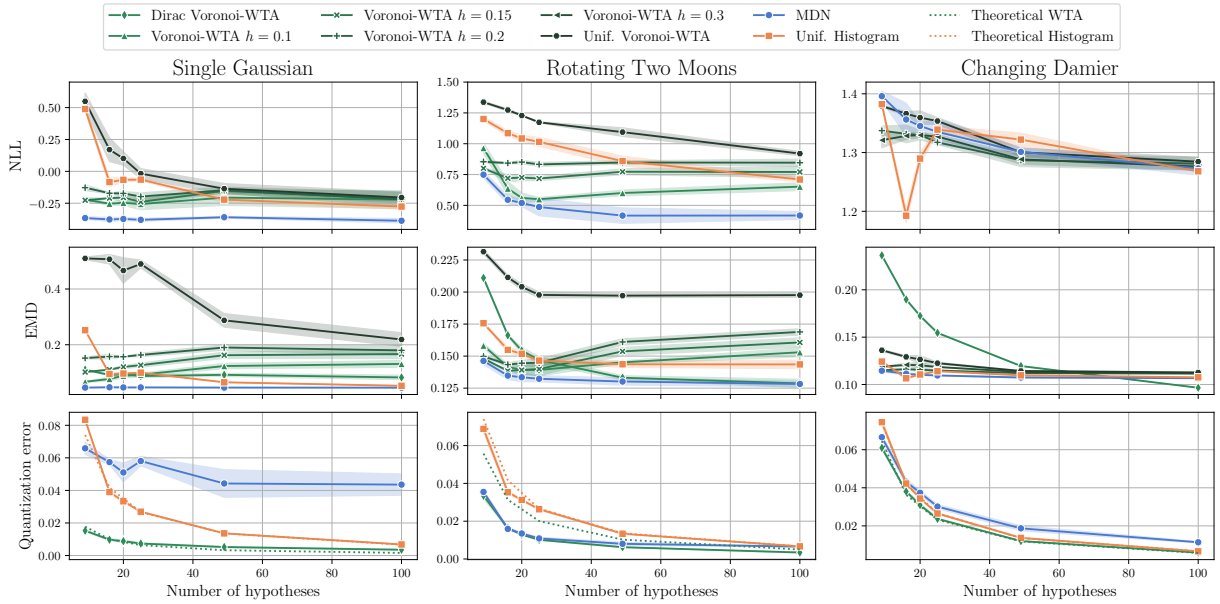

Figure 7. Standard deviations across three random seeds are shown in the results of Figure 3. To simplify the presentation, only the following Voronoi-WTA curves are displayed: $h = 0.1, 0.15, 0.2$ for the Single Gaussian and Rotating Moons datasets, and $h = 0.2, 0.3$ for the Changing Damier dataset. Additionally, the y-axis of the EMD plot for the Changing Damier was cropped in Figure 3 to enhance the scale readability.

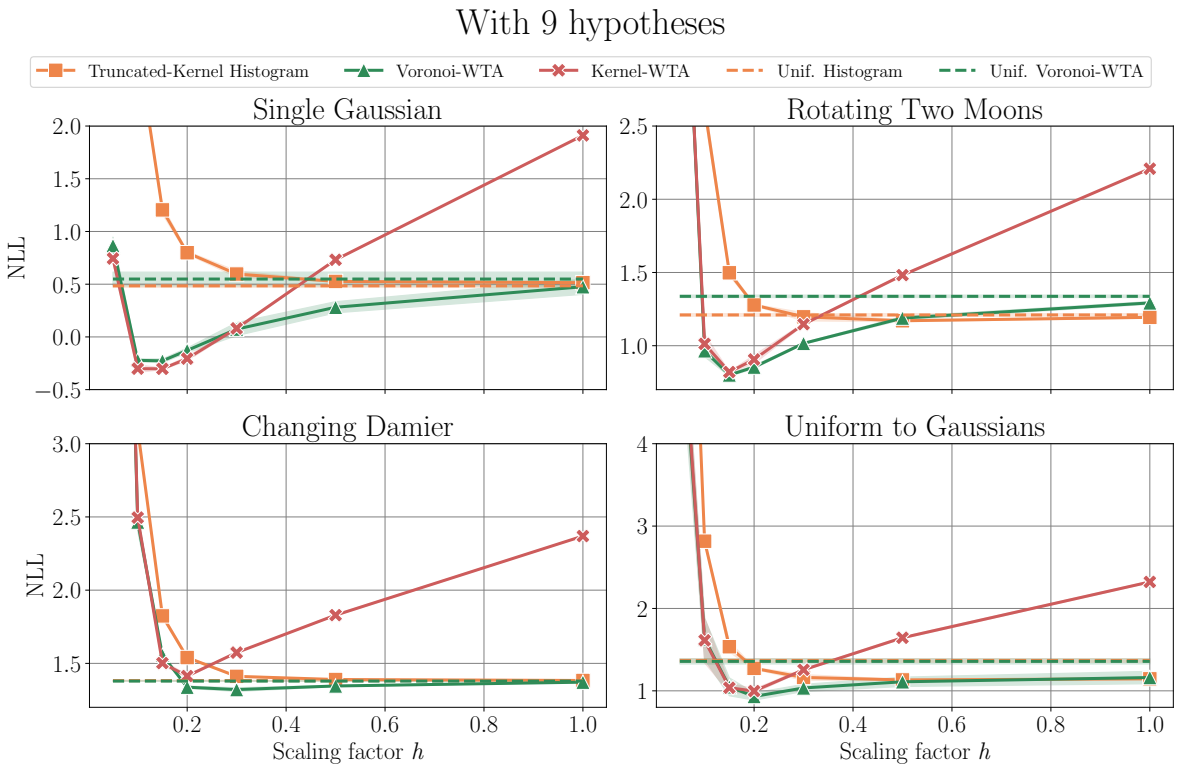

Figure 8. **NLL *vs.* h with 9 hypotheses.**

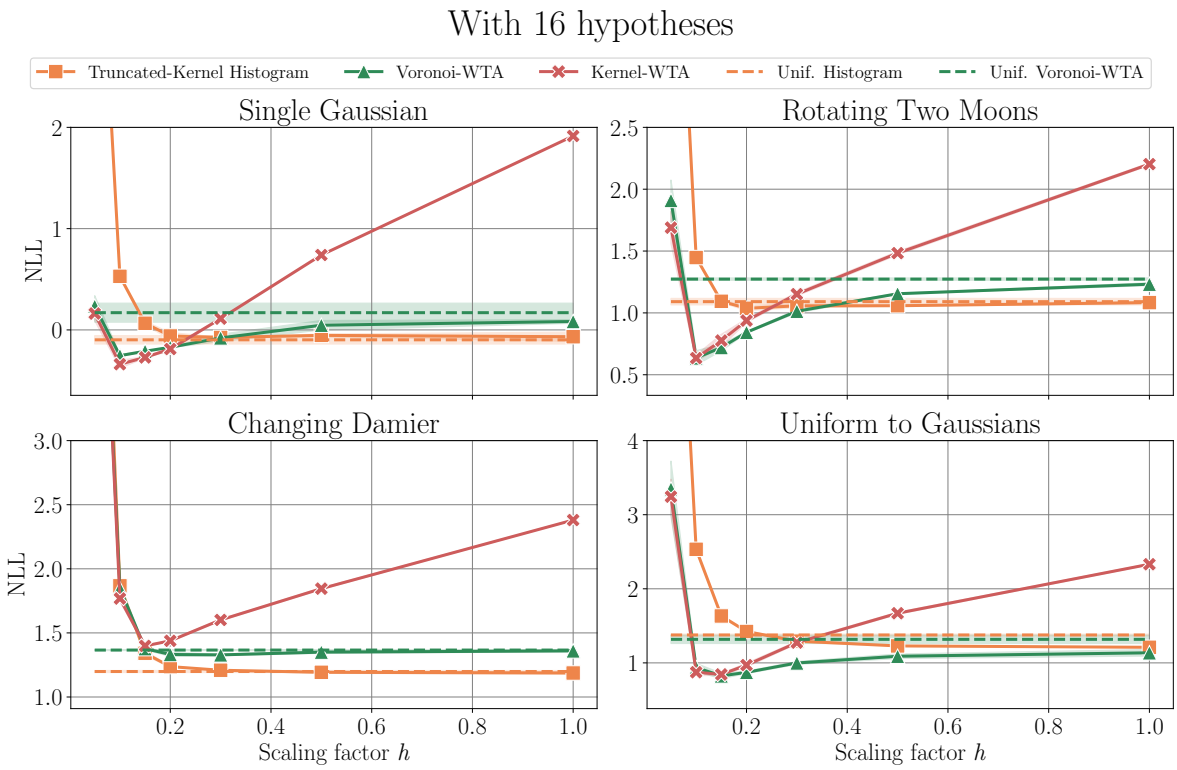

Figure 9. **NLL *vs.* h with 16 hypotheses.**

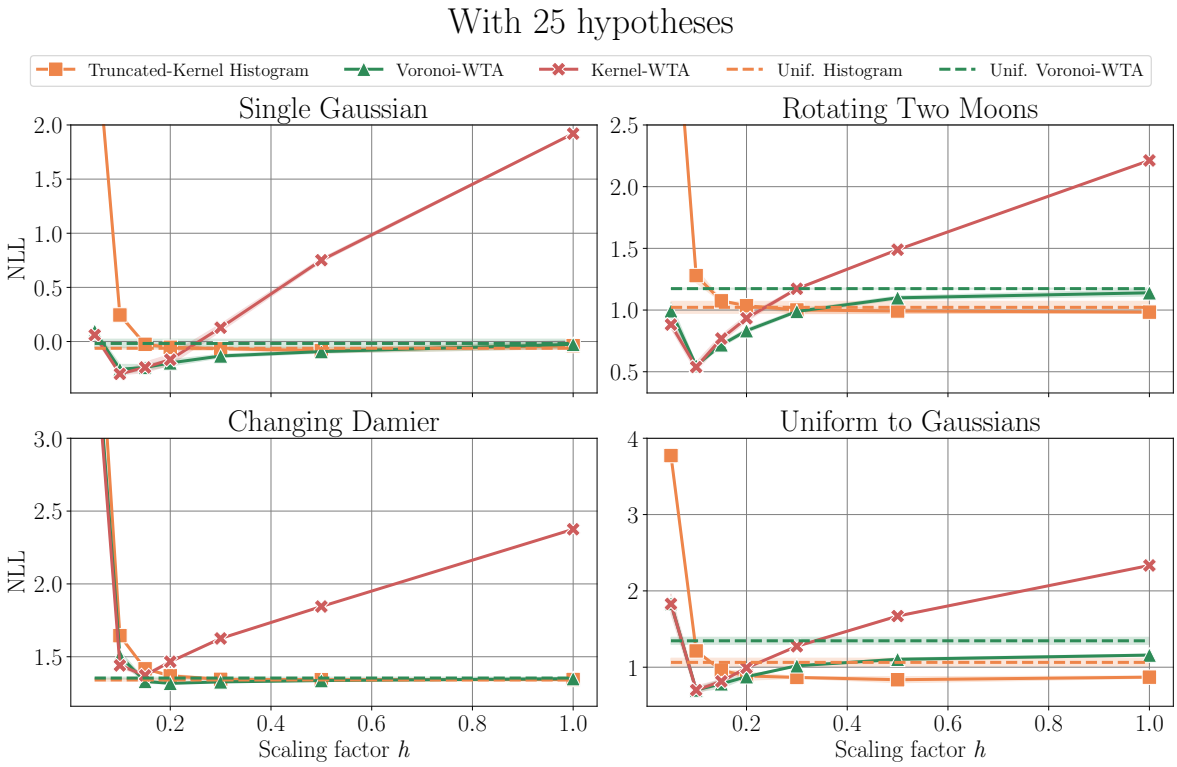

Figure 10. **NLL *vs.* h with 25 hypotheses.**

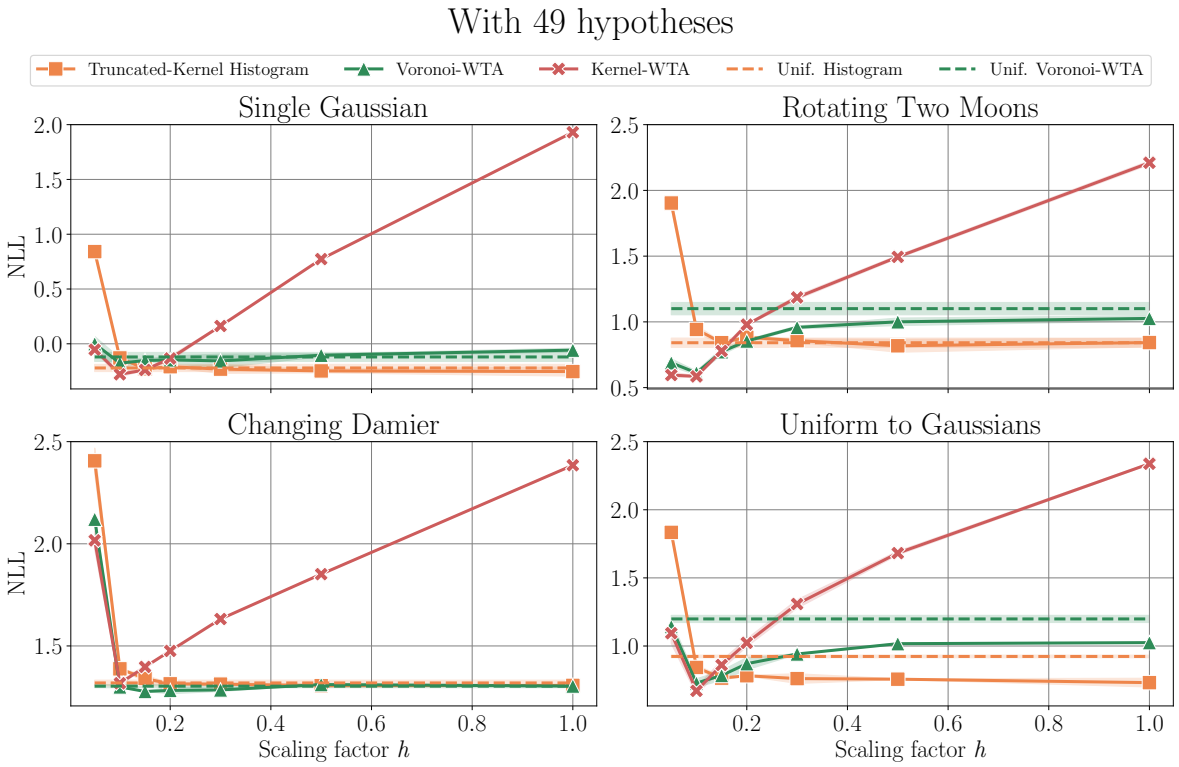

Figure 11. **NLL *vs.* h with 49 hypotheses.**

C.1.5. Additional results

Additional results on the four synthetic datasets *Rotating Two Moons*, *Changing Damier*, *Single Gaussian* and *Uniform to Gaussians* presented in Section 6.1 are provided in Figures 8, 9, 10, 11. As in Figure 4, the aim here is to demonstrate the resilience of Voronoi-WTA with respect to the choice of the scaling factor h in comparison with different baselines. These include Kernel-WTA, and Truncated-Kernel Histogram, where truncated kernels were placed on the bin position. Here, the truncated kernel variants were computed with normalized Gaussian kernels (41) with $d = 2$ here. Uniform kernels were used in the baselines 'Unif. Histogram' and 'Unif. Voronoi-WTA'.

At first, several sanity checks can be carried out. One can verify that as $h \to 0$, Kernel-WTA and Voronoi-WTA are equivalent in all datasets. This outcome is anticipated since, in such regimes, the impact of truncation is virtually negligible, as discussed in Section 4.2. Similarly, when h approaches infinity, both Voronoi-WTA and the Truncated-Kernel Histogram tend to align with their respective uniform versions – Uniform Voronoi-WTA and Uniform Histogram. This behavior is expected due to the bounded nature of the output space. Note that we empirically observe a convex U-shape for the NLL curves on the validation set of Kernel-WTA and Voronoi-WTA as a function of h, allowing us to use adaptive grid search algorithms, such as the Golden Section Search (Kiefer, 1953), which have a fast convergence rate in this setting.

Additionally, observations consistent with those discussed in Section 4 have been made. These findings highlight the robustness of Voronoi-WTA against variations in the h parameter, demonstrating its superior performance over Kernel-WTA, at larger values of h. In contrast, Voronoi-WTA's advantages over the Truncated-Kernel Histogram become more pronounced at smaller h values. Except for the Changing Damier dataset, where the Histogram shows an immediate advantage, the results align with the analysis of Section 6.3, indicating that the Histogram's performance suffers from suboptimal hypothesis placement in this regime. However, it is important to note that as h increases, the Truncated-Kernel Histogram sometimes matches or exceeds the performance of other methods at a fixed h. Likewise, when K is large, the Histogram achieves greater resolution and becomes competitive, thus compensating the naive placement of the hypotheses.

The slight deterioration in performance when considering uniform kernels compared with Gaussian kernels is attributed to the so-called 'compactification' issue , as discussed in Polianskii et al. (2022). Indeed, in scenarios with large cells in the

Voronoi tesselation,[2] the application of a uniform kernel can significantly worsen NLL results. This effect is indeed less pronounced in the histogram method, where the volume of each cell remains constant.

Consistently with observations made from audio data discussed in Section 6.5, Figures 8, 9, 10, and 11 illustrate that the performance gap between Kernel-WTA and Voronoi-WTA widens as h increases and the number of hypotheses grows. It is important to acknowledge a limitation: in scenarios where the hypotheses are few and sufficiently spaced apart, the impact of truncation becomes minor. Under these conditions, Kernel-WTA is likely to offer comparable performance to Voronoi-WTA for a wide range of h values.

C.2. UCI datasets

We conducted additional experiments on the UCI Regression Datasets (Dua & Graff, 2017), which are a standard benchmark to evaluate conditional density estimators. Table 3 provides the sizes of the datasets.

Experimental setup. All the estimators mentioned in our manuscript (Mixture Density Network, Histogram-based methods, and WTA-based methods) are trained and evaluated on these datasets, and results are provided in Table 4 and Table 5. The results were computed following the same experimental protocol from Hernández-Lobato & Adams (2015). In particular, each dataset is divided into 20 train-test folds, except the protein dataset, which is divided into 5 folds, and the Year Prediction MSD dataset for which a single train-test split is used. Moreover, we use the same neural network backbone for each baseline: a one-hidden layer MLP with ReLU activation function, containing 50 hidden units except for the Protein and Year datasets, for which 100 hidden units were utilized. Each model was trained using the Adam optimizer over $1,000$ epochs with a constant learning rate of 0.01. Our data loading pipeline for the UCI datasets was adapted from the open-sourced implementation of Han et al. (2022). In the results presented in Table 4 and Table 5, we follow the convention of highlighting the best models for each dataset in bold, based on the mean value of the metrics. Additionally, any model whose confidence interval overlaps with this best mean is also bolded.

Baselines. Those tables include results from three baselines reported from Table 1 of Lakshminarayanan et al. (2017) which we use as references for those benchmarks: 'PBP' stands for Probabilistic Back Propagation (Hernández-Lobato & Adams, 2015), and 'MC-dropout' corresponds to Monte Carlo Dropout (Gal & Ghahramani, 2016). The Histogram NLL was computed with truncated kernels (TK-NLL for Truncated-Kernel-Histogram), following the same tuning protocol for h as Voronoi-WTA (V-WTA) and Kernel-WTA (K-WTA). MDN corresponds to a mixture density network with Gaussian kernels as in Appendix C.1.1. The multi-hypotheses baselines (MDN, Histogram, and WTA-based methods) were trained with $K = 5$. In this setup, the regular grid of the histogram was defined with $f_\theta^k(x) = \frac{2(k-3)}{5}$ for $k \in \{1, \dots, 5\}$.

Metrics. The computed metrics correspond to first the RMSE, which is defined as $\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i \ell(\hat{y}_i, y_i)}$, where $\hat{y}_i$ denotes the estimated conditional mean, which was estimated with $\sum_{k=1}^K \gamma_k(x) z_k(x)$ for the WTA variants, and $\sum_{k=1}^K \pi_k(x)\mu_k(x)$ for MDN. NLL has been calculated in the same way as in the previous sections, with $\dim(\mathcal{Y}) = 1$.

Evaluation details. Our density estimators were trained according to the procedures outlined in Section 6. Post-training, the scaling factor h was tuned based on the average NLL over the validation set (20 % of the training data) using a golden section search (Kiefer, 1953) (with tolerance set to 0.1 and the search interval bounded by $[0.1, 2]$), following a similar protocol as Gal & Ghahramani (2016). In this setup, as per the guidelines in Hernández-Lobato & Adams (2015), we normalized both input and output variables for training using the means and standard deviations from the training data. For evaluation, we restored the original scale of the output predictions with the transformation $f_\theta^k(x) \mapsto \mu_{\text{train}} + \sigma_{\text{train}} f_\theta^k(x)$ where μ_{train} and σ_{train} represent the empirical mean and standard deviation of the response variable across the training set. This transformation also applies to the predicted means of the MDN, while the predicted standard deviations were simply multiplied by μ_{train}.

Results confirm the claims made in our manuscript and provide the following insights:

- Voronoi-WTA outperforms the Histogram-based estimator in terms of NLL and performs comparably to Kernel-WTA. This aligns with the findings discussed in C.1.5, where Voronoi-WTA's edge over Kernel-WTA is less pronounced in settings with a limited number of hypotheses and where h has been already optimized. Future research will examine how accurately optimizing h during validation affects performance when there is a distribution shift between validation and test samples. Such a study could specifically assess potential performance discrepancies between Voronoi-WTA

[2] See for instance the hypotheses placement of a WTA-based trained model on a Gaussian distribution in Figure 2

Table 3. **UCI Regression benchmark datasets.** N is the number of data samples and $\dim(\mathcal{X})$ the input dimension. Here, the output space $\mathcal{Y}$ is one dimensional.

Dataset	Boston	Concrete	Energy	Kin8nm	Naval	Power	Protein	Wine	Yacht	Year
$(N; \dim(\mathcal{X}))$	(506;13)	(1030;8)	(768;8)	(8192;14)	(11934;16)	(9568;4)	(45730;9)	(1599;11)	(308;6)	(515345;90)

and Kernel-WTA under these conditions.

- WTA-based estimators outperform the Histogram-based estimator in terms of RMSE.
- Voronoi-WTA is competitive with MDN for NLL. This is a promising result as the NLL is optimized only during validation for Voronoi-WTA, and during training for MDN. Moreover, we faced stability issues when training MDN (*e.g.,* numerical overflows in log-likelihood computation), that we did not encounter with Voronoi-WTA.

Looking at Table 4 and Table 5, we can see that for data-intensive tasks (Protein, Year) Voronoi-WTA is on par with standard baselines in those benchmarks, and occasionally outperforms them, both in terms of NLL and RMSE. However, for tasks with limited data available, it seems that WTA is not the most suitable method. This underperformance in the small data regime could be expected: because of the competitive nature of the WTA training scheme, each prediction model only sees a fraction of the data.

Table 4. **UCI regression benchmark datasets comparing NLL with 5 hypotheses.** * corresponds to reported results from Lakshminarayanan et al. (2017). '–' corresponds to cases where MDN has not converged. Best results are in **bold**. $\pm$ represents the standard deviation over the splits (non-applicable for the year dataset).

	NLL (↓)						
Datasets	PBP*	MC Dropout*	Deep Ensembles*	MDN	TK-Hist	K-WTA	V-WTA
Boston	**2.57 ± 0.09**	**2.46 ± 0.25**	**2.41 ± 0.25**	2.95 ± 0.31	2.83 ± 0.17	**2.48 ± 0.16**	**2.48 ± 0.19**
Concrete	3.16 ± 0.02	**3.04 ± 0.09**	**3.06 ± 0.18**	3.96 ± 0.24	3.47 ± 0.12	**3.09 ± 0.10**	**3.08 ± 0.12**
Energy	2.04 ± 0.02	1.99 ± 0.09	**1.38 ± 0.22**	**1.25 ± 0.25**	2.39 ± 0.11	2.27 ± 1.22	2.22 ± 1.20
Kin8nm	-0.90 ± 0.01	-0.95 ± 0.03	**-1.20 ± 0.02**	-0.87 ± 0.05	-0.70 ± 0.03	-0.73 ± 0.03	-0.85 ± 0.05
Naval	-3.73 ± 0.01	-3.80 ± 0.05	**-5.63 ± 0.05**	**-5.47 ± 0.29**	-3.06 ± 0.02	-1.94 ± 0.00	-3.52 ± 0.38
Power	2.84 ± 0.01	**2.80 ± 0.05**	**2.79 ± 0.04**	3.02 ± 0.07	3.18 ± 0.02	**2.81 ± 0.05**	**2.85 ± 0.06**
Protein	2.97 ± 0.00	2.89 ± 0.01	2.83 ± 0.02	–	2.64 ± 0.01	**2.39 ± 0.03**	**2.42 ± 0.04**
Wine	0.97 ± 0.01	0.93 ± 0.06	0.94 ± 0.12	**-1.53 ± 0.76**	0.46 ± 0.10	0.42 ± 0.18	0.37 ± 0.17
Yacht	1.63 ± 0.02	1.55 ± 0.12	**1.18 ± 0.21**	2.43 ± 0.72	2.80 ± 0.23	2.23 ± 0.52	2.05 ± 0.46
Year	3.60 ± NA	3.59 ± NA	3.35 ± NA	–	3.57 ± NA	**3.26 ± NA**	3.29 ± NA

Table 5. **UCI regression benchmark datasets comparing RMSE with 5 hypotheses.** * corresponds to reported results from Lakshminarayanan et al. (2017). '–' corresponds to cases where MDN has not converged. Best results are in **bold**. $\pm$ represents the standard deviation over the splits (non-applicable for the year dataset).

	RMSE (↓)						
Datasets	PBP*	MC Dropout*	Deep Ensembles*	MDN	TK-Hist	K-WTA	V-WTA
Boston	**3.01 ± 0.18**	**2.97 ± 0.85**	**3.28 ± 1.00**	**3.65 ± 1.15**	5.53 ± 1.26	**3.54 ± 1.16**	**3.54 ± 1.16**
Concrete	**5.67 ± 0.09**	**5.23 ± 0.53**	6.03 ± 0.58	7.52 ± 0.96	9.03 ± 0.68	6.02 ± 0.65	6.02 ± 0.65
Energy	1.80 ± 0.05	**1.66 ± 0.19**	2.09 ± 0.29	2.35 ± 0.45	3.89 ± 0.48	2.53 ± 0.99	2.53 ± 0.99
Kin8nm	0.10 ± 0.00	0.10 ± 0.00	0.09 ± 0.00	**0.08 ± 0.00**	0.14 ± 0.01	0.10 ± 0.01	0.10 ± 0.01
Naval	0.01 ± 0.00	0.01 ± 0.00	**0.00 ± 0.00**	**0.00 ± 0.00**	0.01 ± 0.00	**0.00 ± 0.00**	**0.00 ± 0.00**
Power	**4.12 ± 0.03**	**4.02 ± 0.18**	**4.11 ± 0.17**	**4.11 ± 0.18**	7.55 ± 0.17	**4.18 ± 0.16**	**4.18 ± 0.16**
Protein	4.73 ± 0.01	**4.36 ± 0.04**	4.71 ± 0.06	–	4.47 ± 0.02	**4.39 ± 0.10**	**4.39 ± 0.10**
Wine	**0.64 ± 0.01**	**0.62 ± 0.04**	**0.64 ± 0.04**	**0.65 ± 0.04**	0.67 ± 0.04	**0.63 ± 0.04**	**0.63 ± 0.04**
Yacht	**1.02 ± 0.05**	**1.11 ± 0.38**	1.58 ± 0.48	4.08 ± 1.57	8.27 ± 2.83	3.28 ± 1.39	3.28 ± 1.39
Year	8.88 ± NA	**8.85 ± NA**	8.89 ± NA	–	9.31 ± NA	9.09 ± NA	9.09 ± NA

C.3. Audio data experiments

C.3.1. Setup

This section describes in greater detail the experimental setup from Section 6.5. The audio data experiments are based on the protocol of Schymura et al. (2021a); Letzelter et al. (2023) which is given as follows. Nevertheless, distinctions are to be made with respect to previous works. While Letzelter et al. (2023) study the case of punctual sound source localization (Grumiaux et al., 2022), we apply our estimators here the more general problem of data uncertainty quantification, due for instance to an actual spatial dispersion of sound sources or annotation errors. This extension is described in greater detail in the paragraph 'Synthetic perturbations' given below.

Dataset preprocessing. In our experiments, we used the ANSYN dataset (Adavanne et al., 2018a), which contains spatially localized sound events under anechoic conditions. We conformed to the dataset processing techniques as detailed in works by Schymura et al. (2021a); Letzelter et al. (2023). We employed the first-order Ambisonics format with four input audio channels. The audio recordings, with a 44.1 kHz sampling rate, were segmented into 30-second durations. These segments were further divided into non-overlapping chunks of 2 s to serve as the basis for training data. Spectrograms were calculated with a Hann window of 0.04 seconds for the Short Term Fourier Transform calculations. This was done with a 50% overlap between frames and utilizing 2048 points for the Fast Fourier Transform computation. The information input into the models included both the amplitude and phase data, stacked channel-wise.

Architecture. We utilized SeldNet (Adavanne et al., 2018a) as backbone (with ~ 1.6 M parameters). The data processing starts with the preprocessing of raw audio, which is then fed into the model in the form of spectrograms of a set duration, including phase information. The model then provides localization outputs at the specified resolution, in this case, considering $T = 100$ output time steps for each segment. The architecture processes the data through feature extraction modules, including Convolutional Neural Networks (CNNs) and Bi-directional Gated Recurrent Unit (GRU) layers, creating a representation for each time step at the determined output resolution. These intermediate representations are subsequently connected to the final localization predictions via Fully Connected (FC) layers. To suit the Winner-takes-all framework, the terminal FC layers are divided into K separate FC heads, each delivering a two-dimensional output (azimuth and elevation) at each time step. Additionally, the system incorporates score heads at the final stage, each yielding a single value between 0 and 1, achieved through a sigmoid activation function. Note that the 'Histogram' baseline in the audio experiments of Section 6.5 utilizes the same backbone with fixed hypotheses heads. More precisely, denoting k the hypothesis index associated with the histogram grid for row $i \in [\![1, N_{\text{rows}}]\!]$ and column $j \in [\![1, N_{\text{cols}}]\!]$ we have, for every $x \in \mathcal{X}$:

$$f_\theta^k(x) = \left(-\pi + \left(i - \frac{1}{2}\right)\frac{2\pi}{N_{\text{rows}}}, -\frac{\pi}{2} + \left(j - \frac{1}{2}\right)\frac{\pi}{N_{\text{cols}}}\right),$$

where these coordinates correspond, respectively, to the azimuth and the elevation within the ranges $[-\pi, \pi]$ and $[-\frac{\pi}{2}, \frac{\pi}{2}]$. Note that for the 2D Histogram, $K = N_{\text{rows}} N_{\text{cols}}$ with the notations of Table 2. The results of for the 2D Histogram were computed with $N_{\text{rows}} = N_{\text{cols}}$, except when $K = 20$ (in Figure 12 and Table 6) where we set $N_{\text{rows}} = 5$ and $N_{\text{cols}} = 4$.

Training details. The trainings were conducted using the AdamW optimizer (Loshchilov & Hutter, 2018), with a batch size of 32, an initial learning rate of 0.05, and following the scheduling scheme from Vaswani et al. (2017). The WTA model was trained using the multi-target version of the Winner-takes-all loss (Equation 2 and 5 of Letzelter et al. (2023)), using confidence weight $\beta = 1$. Note that as the predictions and the targets belong to the unit sphere, the underlying loss ℓ used was the spherical distance $\ell(\hat{y}, y) = \arccos[\hat{y}^\top y]$ where $y, \hat{y} \in \mathbb{S}^2 \subseteq \mathbb{R}^3$.

Synthetic perturbation. To properly evaluate the ability of the baseline methods to predict conditional distributions, we propose a protocol that involves injecting heteroscedastic noise.

1. During training, we performed *class-conditioned perturbation* of the target source positions. More precisely, for each of the $n_{\text{c}} = 11$ classes in the dataset (`speech, phone, keyboard, doorslam, laughter, keysDrop, pageturn, drawer, cough, clearthroat, knock`), we randomly perturbed the target position of the sound sources using a distinct standard deviation assigned to each class. These perturbations were drawn from normal distributions on the spherical coordinates, with angular standard deviations (in degrees) in the set $\{5 + 5c \mid c \in [\![0, n_{\text{c}} - 1]\!]\}$.
2. Once trained, the hope is that at inference time, the models have understood the ambiguity in the data, such that they can infer, given a new input audio snippet, the spatial spread of the sound source.
3. We evaluate the quality of the estimated distribution with regard to the ground-truth distribution, which is known here.

Evaluation details. For preserving the geometry of the sphere, the likelihoods of Kernel-WTA and Voronoi-WTA were computed using isotropic von Mises-Fisher kernels. For data that lives on a 2-dimensional sphere, *i.e.,* for $f_\theta^k(x), y \in \mathbb{S}^2 \subseteq \mathbb{R}^3$, such kernels write in the form

$$K_h(f_\theta^k(x), y) = \frac{\kappa}{4\pi \sinh \kappa} \exp\left(\kappa y^\top f_\theta^k(x)\right) ,$$

where $h > 0$ is the scaling factor associated with the kernel defined as $h = \frac{1}{\sqrt{\kappa}}$. The kernel is thus defined with a single concentration parameter $\kappa > 0$. The volume $V(f_\theta^k(x), K_h) \triangleq \int_{\mathcal{Y}_\theta^k(x)} K_h\left(f_\theta^k(x), \tilde{y}\right) \mathrm{d}\tilde{y}$ of the predicted distribution of Voronoi-WTA in each curved cell k was computed noting that

$$\int_{\mathcal{Y}_\theta^k(x)} K_h\left(f_\theta^k(x), \tilde{y}\right) \mathrm{d}\tilde{y} \simeq \mathcal{S}(\mathcal{Y}_\theta^k(x)) \mathbb{E}_{Y \sim \mathcal{U}(\mathcal{Y}_\theta^k(x))}[K_h(f_\theta^k(x), Y)] , \tag{44}$$

where $\mathcal{S}(\mathcal{Y}_\theta^k(x))$ is the surface of cell k. Note that we also have

$$\mathcal{S}(\mathcal{Y}_\theta^k(x)) = 4\pi \mathbb{E}_{Y \sim \mathcal{U}(\mathbb{S}^2)}[\mathbb{1}(Y \in \mathcal{Y}_\theta^k(x))] . \tag{45}$$

In practice, the expectations in (44) and (45) were computed using monte-carlo estimates, *i.e.*,

$$\mathbb{E}_{Y \sim \mathcal{U}(\mathcal{Y}_\theta^k(x))}[K_h(f_\theta^k(x), Y)] \simeq \frac{1}{N'} \sum_{y_i \sim \mathcal{U}(\mathcal{Y}_\theta^k(x))} K_h(f_\theta^k(x), y_i) ,$$

$$\mathbb{E}_{Y \sim \mathcal{U}(\mathbb{S}^2)}[\mathbb{1}(Y \in \mathcal{Y}_\theta^k(x))] \simeq \frac{1}{N'} \sum_{y_i \sim \mathcal{U}(\mathbb{S}^2)} \mathbb{1}(y_i \in \mathcal{Y}_\theta^k(x)) ,$$

where N' is the number of sampled points. The uniform sampling on the unit sphere $\mathcal{U}(\mathbb{S}^2)$ has been performed by sampling the azimuth and elevation angles with $\phi \sim \mathcal{U}([0, 2\pi])$ and $\theta \sim \arccos[\mathcal{U}([-1, 1])] - \pi/2$ (Weisstein, 2002).

The NLL defined in (35) was itself computed using a single sample from the target distribution ρ_x for each input, which can be assumed to be known in the synthetic data perturbation setup (which is as a mixture of Gaussians here; see Figure 13 for an illustration).

For adapting the quantization error to the spherical geometry, (14) was generalized with

$$\mathcal{R}(\mathcal{Z}) = \int_{\mathcal{Y}} \min_{z \in \mathcal{Z}} \mathrm{dist}(y, z)^2 \, \rho_x(y) \mathrm{d}y , \tag{46}$$

where dist is the spherical distance defined as $\mathrm{dist}(y, z) = \arccos[y^\top z]$ for $y, z \in \mathbb{S}^2 \subseteq \mathbb{R}^3$.

Our implementation was based on Schymura et al. (2021a;b); Letzelter et al. (2023); Polianskii et al. (2022).

C.3.2. Additional results

Additional results complementing those in Table 2 in the ANSYN audio dataset are presented in Figure 12 and in Table 6, corresponding to $9, 16, 20$ and 25 hypotheses, respectively.

In Figure 12 each subplot illustrates the Negative Log-Likelihood (NLL) on the test set as a function of the scaling factor h, across different baselines. The legends in the figures follow the format of Figure 4: 'WTA' denotes methods trained using the Winner-takes-all scheme, 'Histogram' refers to the Histogram baseline, 'Voronoi' denotes the application of truncated kernels based on the 'Histogram' or 'WTA' hypotheses, and 'Kernel-WTA' is defined in Section 4.1. 'Unif.' indicates the utilization of uniform kernels instead of von Mises-Fisher kernels, when using truncated kernels estimators. Here, we introduce an additional baseline for further discussion: a mixture density network based on the von Mises-Fisher distribution (M-vMF). It is based on the same training loss as (33), but considering instead a mixture of von Mises-Fisher distribution

$$\hat{\rho}_\theta(y|x) = \sum_{k=1}^{K} \pi_k(x) \frac{\kappa_k(x)}{4\pi \sinh \kappa_k(x)} \exp\left(\kappa_k(x) y^\top \mu_k(x)\right) ,$$

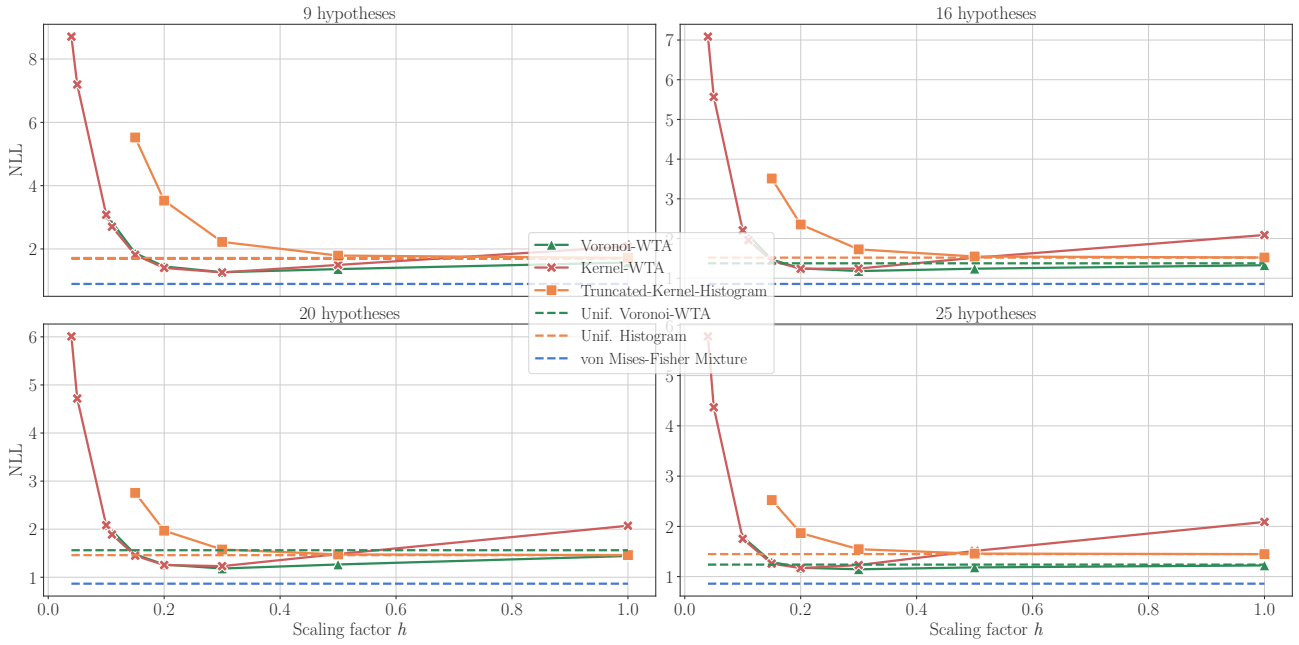

Figure 12. NLL ***vs.*** h **results on ANSYN with 9, 16, 20 and 25 hypotheses.** NLL on the test set of the spatial audio dataset ANSYN (see Section 6.5) as a function of the scaling factor h. 'Von Mises mixture' corresponds to a von Mises-Fisher-based mixture density network. The legend is the same as in Figure 4. 'Truncated-Histogram' corresponds to the standard Histogram where truncated kernels are placed on the fixed hypotheses, instead of standard uniform ones. 'Unif. Histogram' and 'Unif. Voronoi WTA' corresponds to uniform truncated kernels, which correspond to the limit of their von Mises-Fisher truncated variant as $h \to \infty$. As expected, 'Kernel-WTA' and 'Voronoi-WTA' coincide as $h \to 0$ (see Section 4.2). Here, dashed lines correspond to baselines that are independent of h.

with parameters $\{\pi_k(x), \mu_k(x), \kappa_k(x)\}$. Stability issues were observed in M-vMF training, including numerical overflow in NLL computation. Those were partially reduced, for instance using the following expression, $\log(\sinh(t)) = t + \log(1 - e^{-2t}) - \log(2)$ allowing computation stability for large $t \in \mathbb{R}_+^*$.

NLL Comparison. First, it is important to highlight the consistency observed in the results as the scaling factor h approaches zero. In this limit, the NLL values for both Voronoi-WTA and Kernel-WTA tend to coincide, an outcome that aligns with expectations; the impact of truncation diminishes in this scenario, as elaborated in Section 6.3. Analogously to observations made in Figure 4, the NLL values associated with the Histogram baseline (orange squares) demonstrate the fastest divergence as $h \to 0$, due to the non-optimal positioning of the hypotheses.

As the scaling factor h increases, a quantitative improvement is observed in the NLL performance of Voronoi-WTA (green triangles) compared to other baseline methods. Notably, we notice that the Kernel Density Estimation (KDE) approaches, namely Kernel-WTA and Unweighted Kernel-WTA, exhibit a performance decline with increasing h, in contrast to the truncated kernel versions (Voronoi-WTA and Voronoi-Histogram, with Indian red color in the plots). This divergence is attributed to the dispersion of probability mass beyond the boundaries of the Voronoi cells in KDE-variants, leading to a loss of local geometric properties, as detailed in Section 4.1.

Furthermore, as outlined in Section 6.5, note that the performance gap between Kernel-WTA and Voronoi-WTA tends to narrow with a decrease in the number of hypotheses. This is because, in scenarios with fewer hypotheses, the impact of kernel truncation becomes less significant. We see that in those settings, the Voronoi-WTA almost reaches the performance of the M-vMF mixture density network, which slightly outperforms the other estimators in terms of NLL, with the gap getting closer when the number of hypotheses is large. This is promising, as von Mises Fisher has three advantages in this context: 1) it optimizes NLL during training; 2) the audio dataset targets are perturbed with synthetic Gaussian angular noise, which is similar to the von Mises Fisher kernel; 3) the M-vMF has more parameters than the Voronoi-WTA method, as it allows for a variable concentration parameter in each cell. For a fairer comparison, experiments on real-world data without synthetic perturbations were performed in Appendix C.2, showing that Voronoi-WTA can even outperform MDN-based methods in terms of NLL in some settings.

Table 6. **Quantization Error comparison.** Quantization Error ($\times 10^2$) for the compared estimators on the spatial audio dataset ANSYN (see Section 6.4) with spherical underlying distances (expressed in radians). M-vMF corresponds to a von Mises-Fisher-based mixture density network.

	K			
Estimator	9	16	20	25
V-WTA	**0.42**	**0.26**	**0.27**	**0.17**
K-WTA	**0.42**	**0.26**	**0.27**	**0.17**
M-vMF	0.62	0.40	0.34	0.29
Hist	1.23	0.72	0.55	0.47

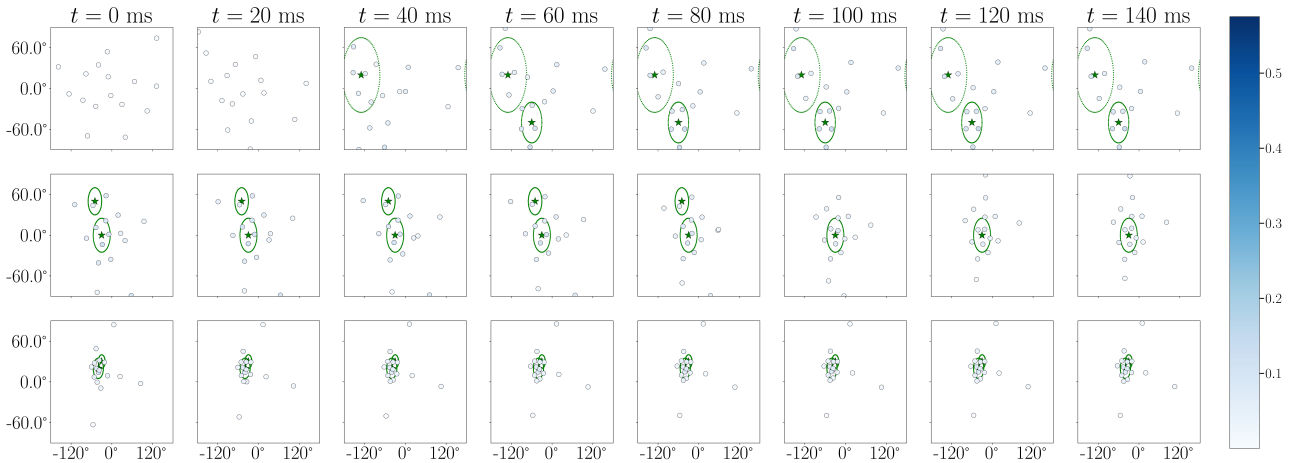

Figure 13. **Quantifying Uncertainty in Spatial Audio with WTA Learners.** Results for audio clips from the test set of the ANSYN dataset. Each row corresponds to distinct recordings, while each column represents different time frames. The subplot's axes correspond to azimuth and elevation in degrees. The target positions of sound sources, marked by green stars, are within ellipses showing their theoretical spatial dispersion. The WTA model's 16 predictions are indicated by blue shaded circles, and the score confidence is indicated by the shade intensity (the legend is given by the color bar on the *right*). We can see that the prediction dispersion follows closely the theoretical dispersion of the sources. This demonstrates that the model successfully grasps input-dependent uncertainty.

Quantization Error comparison. Quantization error results are provided in Table 6, are consistent with the results of Section 6.5. First, we see that for all methods, the quantization improves with K, which is consistent. Secondly, we see an advantage in the quantization error between the WTA-based and the M-vMF methods. This is consistent with the third line of Figure 3, which shows that WTA tends to outperform MDN in terms of quantization error.

Visualizations of the WTA predictions on audio data are also provided in Figure 13.

C.4. Computation details

In this research, we utilized the Python programming language, along with the Pytorch (Paszke et al., 2019) deep learning framework. We also employed the Hydra and MLFlow libraries for experimental purposes. Our coding was inspired by several previous works (Adavanne et al., 2019; Makansi et al., 2019; Schymura et al., 2021a; Polianskii et al., 2022; Han et al., 2022; Letzelter et al., 2023). The training of our neural networks was conducted on NVIDIA A100 GPUs.